`

# Computer Science & Information Technology 150

## Advances in Machine Learning, Data Mining and Computing

`

David C. Wyld,
Dhinaharan Nagamalai (Eds)

# Computer Science & Information Technology

- 2nd International Conference on Machine Learning Techniques and NLP (MLNLP 2021), September 18 ~ 19, 2021, Copenhagen, Denmark
- 7th International Conference on Computer Science, Engineering and Information Technology (CSITY 2021)
- 7th International Conference on Networks & Communications (NWCOM 2021)
- 7th International Conference on Signal and Image Processing (SIGPRO 2021)
- 2nd International Conference on Advances in Software Engineering (ASOFT 2021)
- 7th International Conference on Artificial Intelligence and Fuzzy Logic Systems (AIFZ 2021)
- 2nd International Conference on Big Data & IOT (BDIoT 2021)
- 8th International Conference on Information Technology, Control, Chaos, Modeling and Applications (ITCCMA 2021)
- 2nd International Conference on Cloud Computing, Security and Blockchain (CLSB 2021)
- 7th International Conference on Data Mining (DTMN 2021)
- 2nd International Conference on Big Data & IOT (BDIoT 2021)

**Published By**

**AIRCC Publishing Corporation**

`

## Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-48-0
DOI: 10.5121/csit.2021.111401 - 10.5121/csit.2021.111429

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

`

# Preface

The International Conference on 2nd International Conference on Machine Learning Techniques and NLP (MLNLP 2021), September 18 ~ 19, 2021, Copenhagen, Denmark, 7th International Conference on Computer Science, Engineering and Information Technology (CSITY 2021), 7th International Conference on Networks & Communications (NWCOM 2021), 7th International Conference on Signal and Image Processing (SIGPRO 2021), 2nd International Conference on Advances in Software Engineering (ASOFT 2021), 7th International Conference on Artificial Intelligence and Fuzzy Logic Systems (AIFZ 2021), 2nd International Conference on Big Data & IOT (BDIoT 2021), 8th International Conference on Information Technology, Control, Chaos, Modeling and Applications (ITCCMA 2021), 2nd International Conference on Cloud Computing, Security and Blockchain (CLSB 2021), 7th International Conference on Data Mining (DTMN 2021) and 2nd International Conference on Big Data & IOT (BDIoT 2021) was collocated with International Conference on 2nd International Conference on Machine Learning Techniques and NLP (MLNLP 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The MLNLP 2021, CSITY 2021, NWCOM 2021, SIGPRO 2021, ASOFT 2021, AIFZ 2021, BDIoT 2021, ITCCMA 2021, CLSB 2021 and DTMN 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, MLNLP 2021, CSITY 2021, NWCOM 2021, SIGPRO 2021, ASOFT 2021, AIFZ 2021, BDIoT 2021, ITCCMA 2021, CLSB 2021 and DTMN 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the MLNLP 2021, CSITY 2021, NWCOM 2021, SIGPRO 2021, ASOFT 2021, AIFZ 2021, BDIoT 2021, ITCCMA 2021, CLSB 2021 and DTMN 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

| General Chair | Organization |
|---|---|
| David C. Wyld, | Southeastern Louisiana University, USA |
| Dhinaharan Nagamalai (Eds) | Wireilla Net Solutions, Australia |

## Program Committee Members

| | |
|---|---|
| Abdel-Badeeh M. Salem, | Ain Shams University, Cairo, Egypt |
| Abdelhadi Assir, | Hassan 1st University, Morocco |
| Abdellatif I. Moustafa, | Umm AL-Qura University, Saudi Arabia |
| Abderrahim Siam, | University of Khenchela, Algeria |
| Abderrahmane Ez-zahout, | Mohamed 5 University, Morocco |
| Abdessamad Belangour, | University Hassan II Casablanca Morocco |
| Abdullah, | Adigrat University, Africa |
| Abdulraqeb Alhammadi, | University of Technology Malaysia (UTM), Malaysia |
| Addisson Salazar, | Universitat Politecnica de València, Spain |
| Adeyanju Sosimi, | University of Lagos, Nigeria |
| Adnan Mohammed Hussein, | Northern Technical University, Iraq |
| Adrian Olaru, | University Politehnica of Bucharest, Romania |
| Ahmad A. Saifan, | Yarmouk University, Jordan |
| Ahmed Farouk AbdelGawad, | Faculty of Engineering, Zagazig Univ, Egypt |
| Ahmed Kadhim Hussein, | University of Babylon, Iraq |
| Aishwarya Asesh, | Adobe, USA |
| Akhil Gupta, | Lovely Professional University, India |
| Akhil Jabbar, | Vardhaman College of Engineering, India |
| Alaa Alahmadi, | Imam Abdulrahman Bin Faisal University, Saudi Arabia |
| Alagarsamy K, | Madurai kamaraj university, India |
| Alex Mathew, | Bethany College, USA |
| Alexander Gelbukh, | Instituto Politécnico Nacional, Mexico |
| Ali Abdrhman Mohammed Ukasha, | Sebha University, Libya |
| Ali Alsabbagh, | Ministry of communication, Iraq |
| Alireza Valipour Baboli, | Technical and Vocational University, Iran |
| Álvaro Rocha, | University of Coimbra, Portugal |
| Amal Zouhri, | Sidi Mohammed Ben Abdellah University, Morocco |
| Aman Jatain, | Amity University, Haryana, India |
| Aman Malhotra, | SRM Institute of Science and Technology, India |
| Amando P. Singun Jr, | University of Technology and Applied Sciences, Oman |
| Amari Houda, | Networking & Telecom Engineering, Tunisia |
| Amel Ourici , | University Badji Mokhtar Annaba, Algeria |
| Amit Mishra, | Baze University, Nigeria |
| Amizah Malip, | University of Malaya, Malaysia |
| Ana Leal, | University of Macau, China |
| Anand Nayyar, | Duy Tan University, Vietnam |
| Anandkumar Mani, | SRM Easwari Engineering College, India |
| Anas M.R. Al Sobeh, | Yarmouk University, Jordan |
| Anastasios Doulamis, | National technical University of Athens, Greece |
| Andy Rachman, | Institut Teknologi Adhi Tama Surabaya, Indonesia |
| Anis Ismail, | Lebanese University, Lebanon |
| Anita Yadav, | Harcourt Butler Technical University, India |
| Ankur Singh Bist, | Chief AI Scientist at Signy Advanced Technology, India |
| Anna Soltysik-Piorunkiewicz, | Uniwersytet Ekonomiczny, Poland |

`

| | |
|---|---|
| Anouar Abtoy, | Abdelmalek Essaâdi University, Morocco |
| António Abreu, | ISEL, Portugal |
| Arif Sari, | European University of Lefke, Cyprus |
| Armir Bujaria, | Studi di Padova, Italy |
| Arthi, | SRM Institute of Science and Technology, India |
| Asghar gholamian, | Faculty of Electrical and Computer Engineering, Babol |
| Ashkan Tashk, | SDU, Denmark |
| Ashok Kumar, | University of Louisiana at Lafayette, USA |
| Asif Irshad Khan, | King Abdulaziz University, KSA |
| Assem Abdel Hamied Moussa, | E Commerece Tech Support Systems Manager, Egypt |
| Assia Djenouhat, | University Badji Mokhtar Annaba, Algeria |
| Atik Kulakli, | American University of the Middle East (AUM), Kuwait |
| Attila Kertesz, | University of Szeged, Hungary |
| Atul Garg, | Chitkara University, Punjab |
| Ayad salhieh, | Australian College, Kuwait |
| Ayman M. Abdalla, | Al-Zaytoonah University of Jordan, Jordan |
| Ayman Sadig, | Ahfad University for Women, Sudan |
| Azeddien M. Sllame, | University of Tripoli, Libya |
| Barkat Warda, | University of Constantine, Algeria |
| Beshair Alsiddiq, | Prince Sultan University, Saudi Arabia |
| Bouchra Marzak, | Hassan II University, Morocco |
| Bouhorma Mohammed, | FST of Tangier, Morocco |
| Boukari Nassim, | skikda university, Algeria |
| boukari nassim, | skikda university, algeria |
| Bououden Sofiane, | University Abbes Laghrour Khenchela, Algeria |
| Cagdas Hakan Aladag, | Hacettepe University, Turkey |
| Carla Osthoff, | National Laboratory for Scientific Computing, Brazil |
| Carlos Becker Westphall, | Federal University of Santa Catarina, Brazil |
| Carlos E. Otero, | The University of Virginia's College at Wise, USA |
| Chandrasekar Vuppalapati, | San Jose State University, California |
| Cheman Shaik, | Senior Technology Consultant at Collabera, USA |
| Cheng Siong Chin, | Newcastle University, Singapore |
| Chenglie Hu, | Carroll University, USA |
| Chin-Chen Chang, | Feng Chia University, Taiwan |
| Ching Tan, | School of Computing and Information Systems, Canada |
| Christian Mancas, | Ovidius University, Romania |
| Chuan-Ming Liu, | National Taipei University of Technology, Taiwan |
| Claudia Jacy Barenco Abbas, | University of Brasilia, Brazil |
| Cristian Rodriguez Rivero, | Universiteit van Amsterdam, Netherlands |
| Dadmehr Rahbari, | Tallinn University of Technology, Estonia |
| Dalia Hanna, | Ryerson University, Canada |
| Daming Feng, | Old Dominion University ,USA |
| Daniel Rosa Canedo, | Federal Institute of Goias, Brazil |
| Dário Ferreira, | University of Beira Interior, Portugal |
| Dhanya Jothimani, | Ryerson University, Canada |
| Dharmendra Sharma, | University of Canberra, Australia |
| Dibya Mukhopadhyay, | University of Alabama, USA |
| Dieter Landes, | University of Applied Sciences Coburg, Germany |
| Dimitris Kanellopoulos, | University of Patras, Greece |
| Dinesh Reddy Vemula, | SRM University, India |
| Ding Wang, | Nankai University, China |
| Domenico Calcaterra, | University of Catania, Italy |

`

| | |
|---|---|
| Dongping Tian, | Baoji University of Arts and Sciences, China |
| Dong-yuan Ge, | Guangxi University of Science and Technology, China |
| El Habib Nfaoui, | Sidi Mohamed Ben Abdellah University, Morocco |
| EL Murabet Amina, | Abdelmalek Essaadi University, Morocco |
| Elzbieta Macioszek, | Silesian University of Technology, Poland |
| Emir Kremic, | Federal Institute of Statistics, Herzegovina |
| Endre Pap, | Singidunum University, Serbia |
| Erdal Ozdogan, | Gazi University, Turkey |
| Esmaiel Nojavani, | University Of Isfahan, Iran |
| Eugénia Moreira Bernardino, | Polytechnic Institute of Leiria, Portugal |
| Ez-Zahout Abderrahmane, | Mohamed 5 University, Morocco |
| Fatiha Merazka, | University of Science and Technology, Algeria |
| Felix J. Garcia Clemente, | University of Murcia, Spain |
| Fernando Zacarias Flores, | Universidad Autonoma de Puebla, Mexico |
| Fiza Saher Faizan, | Dhacss beachview Campus , Pakistan |
| Francesco Zirilli, | Sapienza Universita Roma, Italy |
| Froilan D. Mobo, | Philippine Merchant Marine Academy, Philippines |
| Fuad Jamour, | University of California, Riverside, United States |
| G.Rajkumar, | N.M.S.S.Vellaichamy Nadar College, India |
| Gabriella Casalino, | University of Bari, Italy |
| Gang Wang, | University of Connecticut, USA |
| Grigorios N. Beligiannis, | University of Patras - Agrinio Campus, Greece |
| Grzegorz Sierpiński, | Silesian University of Technology, Poland |
| Guilong Liu, | Beijing Language and Culture University, China |
| Gulden Kokturk, | Dokuz Eylül University, Turkey |
| Hala Abukhalaf , | Palestine Polytechnic University, Palestine |
| Hamed Taherdoost, | Hamta Business Corporation, Canada |
| Hamid Ali Abed AL-Asadi, | Iraq University college, Iraq |
| Hamid Khemissa, | USTHB University Algiers, Algeria |
| Hamidreza Rokhsati, | Sapienza University of Rome, Italy |
| Haqi Khalid, | University Putra Malaysia, Malaysia |
| Hedayat Omidvar, | Research & Technology Dept, Iran |
| Hemashree, | Hindusthan College of Arts and Science, India |
| Herbert Kuchen, | University of Münster, Germany |
| Hilal adnan fadhil, | Al -farabi university college, Iraq |
| HlaingHtakeKhaungTin, | University of Computer Studies, Myanmar |
| Hossein Bavarsad, | Mechanical Engineer and Project Manager, Iran |
| Hunyadi Daniel, | Lucian Blaga University of Sibiu, Romania |
| Husam Suleiman, | University of Waterloo, Canada |
| Ilango Velchamy, | CMR Institute of Technology, India |
| Ilham Huseyinov, | Istanbul Aydin University, Turkey |
| Ines Bayoudh Saadi, | Tunis University, Tunisia |
| Isa Maleki, | Islamic Azad University, Iran |
| Israashaker Alani, | Ministry of Science and Technology, Iraq |
| Ivan Izonin, | Lviv Polytechnic National University, Ukraine |
| Iyad Alazzam, | Yarmouk University, Jordan |
| Javier Gozalvez, | Universidad Miguel Hernandez de Elche, Spain |
| Jawad K. Ali, | University of Technology, Iraq |
| Jaymer Jayoma, | Caraga State University, Philippines |
| Jesuk Ko, | Universidad Mayor de San Andres (UMSA), Bolivia |
| Jeyanthi N, | Vellore Institute of Technology, India |
| Jia Ying Ou, | York University, Canada |

`

| | |
|---|---|
| Jian Wang, | China University of Petroleum (East China), China |
| Jianyi Lin, | Khalifa University, United Arab Emirates |
| Jitendra Chauhan, | iViZ Technosolutions Pvt Ltd, India |
| Joao Antonio Aparecido Cardoso, | The Federal Institute of São Paulo, Brazil |
| Jonah Lissner, | technion - israel institute of technology, Israel |
| Jong P. Yoon, | Mercy College-New York, USA |
| Jong-Ha Lee, | Keimyung University, South Korea |
| Jose L. Abellan, | Catholic University of Murcia, Spain |
| Juntao Fei, | Hohai University, P. R. China |
| Kalpesh Wandra, | Gujarat Technological University, India |
| Kamaljit I. Lakhtaria, | Atmiya Institute of Technology & Science-Rajkot, India |
| Kamel Benachenhou, | Blida University, Algeria |
| Kamel Hussein Rahouma, | Minia University, Egypt |
| Kamel Jemai, | University of Gabes, Tunisia |
| Karim El Moutaouakil, | USMBA/FPT of Taza,Morroco |
| Kashif Munir, | University of Hafr Al Batin, Saudi Arabia |
| Kazim Yildiz, | Marmara University, Turkey |
| Kazuyuki Matsumoto, | Tokushima Univesity, Japan |
| Ke-Lin Du, | Concordia University, Canada |
| Khalid M.O Nahar, | Yarmouk University, Jordan |
| Kire Jakimoski, | FON University, Republic of Macedonia |
| Kiril Alexiev, | University of Diyala, Iraq |
| Kirtikumar Patel, | I&E Engineer, USA |
| Klenilmar Lopes Dias, | Federal Institute of Amapa, Brazil |
| Klimis Ntalianis, | Athens University of Applied Sciences, Greece |
| Koh You Beng, | University of Malaya, Malaysia |
| Kolla Bhanu Prakash, | KL University, India |
| Lal Pratap Verma, | Moradabad Institute of Technology Moradabad, India |
| Legand L. Burge, | Howard University, USA |
| Litao Guo, | Xiamen University of Technology,China |
| Luca De Cicco, | Politecnico di Bari, Italy |
| Luigi Patrono, | University of Salento, Italy |
| Luisa Maria Arvide, | Cambra University of Almeria, Spain |
| M. Akhil Jabbar, | Vardhaman College of Engineering, India |
| Maad M. Mijwil, | Baghdad College of Economic Sciences University, Iraq |
| Mabroukah Amarif, | Sebha University, Libya |
| Mahdi Sabri, | Islamic Azad University Urmia Branch, Iran |
| Mahendra Bhatu Gawali, | Sanjivani Group of Institutes, Kopargaon |
| Mahmoud M. Hammad, | Jordan University of Science and Technology, Jordan |
| Mahsa Mohaghegh, | Auckland University of Technology, New Zealand |
| Mai Zaki, | American University of Sharjah, UAE |
| Malka N. Halgamuge, | Melbourne School of Engineering, Australia |
| Mamoun Alazab, | Charles Darwin University, Australia |
| Marcin Paprzycki, | Adam Mickiewicz University, Poland |
| Maria Brojboiu, | University of Craiova, Romania |
| Maumita Bhattacharya, | Charles Sturt University, Australia |
| Md Azher Uddin, | Ajou University, South Korea |
| Mehdi Nezhadnaderi, | Islamic Azad University, Iran |
| Meriem Riahi, | ENSIT, Tunisia |
| Mervat Bamiah, | Alamhj for IT Consultancy, Saudi Arabia |
| Michail Kalogiannakis, | University of Crete, Greece |
| Mihai Carabas, | University POLITEHNICA of Bucharest, Romania |

`

| | |
|---|---|
| Mihai Horia Zaharia, | Gheorghe Asachi Technical University, Romania |
| Ming An Chung, | National Taipei University of Technology, Taiwan |
| Mirsaeid Hosseini Shirvani, | Islamic Azad University, Iran |
| Mohamed Arezki Mellal, | M'Hamed Bougara University, Algeria |
| Mohamed Elhoseny, | American University in the Emirates, UAE |
| Mohamed Ismail Roushdy, | Ain Shams University, Egypt |
| Mohammad A. Alodat, | Sur University College, Oman |
| Mohammad Ashraf Ottom, | Yarmouk University, Jordon |
| Mohammad Siraj, | King Saud University, Saudi Arabia |
| Mohammed Qbadou, | SSDIA Hassan II University of Casablanca, Morocco |
| Morteza Alinia Ahandani, | University of Tabriz, Iran |
| Mostafa EL Mallahi, | Sidi Mohamed Ben Abdellah University, Morocco |
| Mourad Chabane Oussalah, | University of Nantes, France |
| Mudhafar Jalil Jassim Ghrabat, | Huazhong University of Science and Technology, China |
| Muhammad Sajjadur Rahim, | University of Rajshahi, Bangladesh |
| Muneer Masadeh Bani Yassein, | Jordan University of Science and Technology, Jordan |
| Murtaza Cicioglu, | Duzce University, Turkey |
| Mu-Song Chen, | Da-Yeh University, Taiwan |
| MV Ramana Murthy, | Osmania University, India |
| Nadia Abd-Alsabour, | Cairo University, Egypt |
| Nahlah Shatnawi, | Yarmouk University, Jordan |
| Nicolas Durand, | Aix-Marseille University, France |
| Nihar Athreyas, | Spero Devices Inc, USA |
| Nikolai Prokopyev, | Kazan Federal University, Russia |
| Noraziah Ahmad, | University Malaysia Pahang, Malaysia |
| Oliver L. Iliev, | FON University, Republic of MACEDONIA |
| Omid Mahdi Ebadati, | Kharazmi University, Tehran |
| Osamah Ibrahim Khalaf, | Al-Nahrain University, Iraq |
| Osman Toker, | Yıldız Technical University, Turkey |
| Ouided SEKHRI, | Frères Mentouri Constantine 1 University, Algeria |
| P. S. Hiremath, | KLE Technological University, India |
| Pablo Cerro Cañizares, | Autonomous University of Madrid, Spain |
| Pascal Lorenz, | University of Haute Alsace France, France |
| Paulo Trigo, | ISEL, Portugal |
| Pavel Loskot, | ZJU-UIUC Institute, China |
| Petra Perner, | FutureLab Artificial Intelligence IBaI-2, Germany |
| Phuoc Tran-Gia, | University of Wuerzburg, Germany |
| Pi-Chung Hsu, | Shu-Te University, Taiwan |
| Piotr Kulczycki, | AGH University of Science and Technology, Poland |
| Pr Leila Hayet Mouss, | University of Batna 2, Algeria |
| Pr. Smain Femmam, | UHA University, France |
| Preetida Vinayakray Jani, | Sardar Patel Institute of Technology, India |
| Przemyslaw Falkowski-Gilski, | Gdansk University of Technology, Poland |
| Quang Hung Do, | University of Transport Technology, Vietnam |
| R Senthil, | Shinas College of technology, Oman |
| R.Arthi, | SRM Institute of Science and Technology, India |
| R.Sujatha, | VIT University, India |
| Rachid Zagrouba, | Imam Abdulrahman Bin Faisal University, Saudi Arabia |
| Rafik Hamza, | NICT, Japan |
| Rahul Kher, | G H Patel College of Engineering & Technology, India |
| Rajeev Kanth, | University of Turku, Finland |
| Rajeev Kaula, | Missouri State University, USA |

`

| | |
|---|---|
| Rakesh Kumar Mahendran, | Anna University, India |
| Ramadan Elaiess, | University of Benghazi, Libya |
| Ramgopal Kashyap, | Amity University Chhattisgarh, India |
| Richa Purohit, | Y Patil INternational University, India |
| Ritika Agarwal, | Amity University, India |
| Roberts Masillamani M, | Hindustan University, India |
| Rodrigo Pérez Fernández, | Universidad Politécnica de Madrid, Spain |
| Ruofei Shen, | AI researcher - Menlo Park, USA |
| S.Taruna, | JK Lakshmipat University, India |
| Sabina Rossi, | Universita Ca' Foscari Venezia, Italy |
| Sabyasachi Pramanik, | Haldia Institute of Technology, India |
| Sachin Kumar, | Kyungpook National University, South Korea |
| sadaqat ur Rehman, | Namal institute Mianwali, Pakistan |
| Said Agoujil, | Moulay Ismail University, Morocco |
| Saikumar Tara, | CMR Technical Campus, Hyderabad India |
| Samir Kumar Bandyopadhyay, | University of Calcutta, India |
| Samrat Kumar Dey, | Dhaka International University, Bangladesh |
| Sandro Ronaldo Bezerra Oliveira, | Federal University of Pará, Brazil |
| Sebastian FritschM, | IT and CS enthusiast, Germany |
| Sébastien Combéfis, | ECAM Brussels Engineering School, Belgium |
| Seppo Sirkemaa, | University in Turku, Finland |
| Seyedsaeid Mirkamali, | Payame Noor University (PNU), Iran |
| Shahid Ali, | AGI Education Ltd, New Zealand |
| Shahnaz N.Shahbazova, | Azerbaijan Technical University, Azerbaijan |
| Shahram Babaie, | Islamic Azad University, Iran |
| Shashikant Patil, | SVKMs NMIMS, India |
| Shing-Tai Pan, | National University of Kaohsiung, Taiwan |
| Shin-Jer Yang, | Soochow University, Taiwan |
| Shoeib Faraj, | Technical And Vocational University Of Urmia, Tehran |
| Shraf Elnagar, | University of Sharjah, UAE |
| Shridhar B Devamane, | K.L.E. Institute of Technology, India |
| Siarry Patrick, | Universite Paris-Est Creteil, France |
| Siddhartha Bhattacharyya, | Christ University, India |
| Simanta Shekhar Sarmah, | Alpha Clinical Systems, USA |
| Smain Femmam, | UHA University France, France |
| Soha rawas, | Beirut Arab University, Lebanon |
| Solley Thomas, | Carmel College for Women, Goa, India |
| Soraya Sedkaoui, | University of Khemis Miliana, Algeria |
| Souhila Silmi, | USTHB university, Algeria |
| Stefano Michieletto, | University of Padova, Italy |
| Subarna Shakya, | Tribhuvan University, Nepal |
| Subhendu Kumar Pani, | Professor, Biju Patnaik University of Technology, India |
| sukhdeep kaur, | punjab technical university, India |
| Sun-yuan Hsieh, | National Cheng Kung University, Taiwan |
| Surabhi Srivastava, | University of KwaZulu-Natal, South Africa |
| Surender Redhu, | Indian Institute of Technology Kanpur, India |
| T. G. Vasista, | International School of Technology and Sciences, India |
| T. Ramayah, | Universiti Sains Malaysia, Malaysia |
| Taha Mohammed Hasan, | University of Diyala, Iraq |
| Taleb zouggar souad, | Oran 2 university, Algeria |
| Tanzila Saba, | Prince Sultan University, Saudi Arabia |
| Tariq Tawfeeq Yousif Alabdullah, | Universiti Sains Malaysia, Malaysia |

`

| | |
|---|---|
| Tatiana Tambouratzis, | University of Piraeus, Greece |
| Tefo Sekgweleo, | Cape Peninsula University of Technology |
| Thaweesak Yingthawornsuk, | King Mongkuts University of Technology, Thailand |
| Titas De, | Senior Data Scientist, India |
| Tri Kurniawan Wijaya, | Technische Universitat Dresden, Germany |
| Tse Guan Tan, | Universiti Malaysia Kelantan, Malaysia |
| Usman Naseem, | University of Sydney, Australia |
| V.Ilango, | CMR Institute of Technology, India |
| Vahideh Hayyolalam, | Koş University, Turkey |
| Valentina E. Balas, | Aurel Vlaicu University of Arad, Romania |
| Valerianus Hashiyana, | University of Namibia, Namibia |
| Venkata Duvvuri, | Oracle Corp & Purdue University, USA |
| Venkata, | Purdue University, USA |
| Vilem Novak, | University of Ostrava, Czech Republic |
| Vinay S, | PES College of Engineering - Mandya, India |
| Viranjay M. Srivastava, | University of KwaZulu-Natal, South Africa |
| Vu Nguyen Huynh Anh, | Université catholique de Louvain, Belgium |
| Wajid Hassan, | Indiana State University, USA |
| Waleed Bin Owais, | Qatar University, Qatar |
| Wei Cai, | Qualcomm technology, USA |
| Wei lu, | Early Warning Academy, China |
| Wenwu Wang, | University of Surrey, UK |
| Wenyuan Zhang, | Tianjin university, China |
| William R Simpson, | Institute for Defense Analyses, USA |
| WU Yung Gi, | Chang Jung Christian University, Taiwan |
| Xiaoye Liu, | University of Southern Queensland, Australia |
| Xiao-Zhi Gao, | University of Eastern Finland, Finland |
| Yakoop R. H. Qasim, | Mechatronics and Robotics Engineering, Yemen |
| Yakoop Razzaz Hamoud Qasim, | Taiz University, Yemen |
| Yasser M. Alginahi, | University of Windsor, Ontario, Canada |
| Yousef Farhaoui, | Moulay Ismail University, MOROCCO |
| Yousef J. Al-Houmaily, | Institute of Public Administration, Saudi Arabia |
| Youssef Taher, | Center of Guidance and Planning Rabat, Morocco |
| Yuan Tian, | Nanjing Institute of Technology, China |
| Yuchen Zheng, | Shihezi University,China |
| Yung Gi Wu, | Chang Jung Christian University, Taiwan |
| Yuping Yan, | Eotvos Loránd Tudományegyetem, Hungary |
| Yuriy Syerov, | Lviv Polytechnic National University, Ukraine |
| Zainab S. Attarbashi, | Universiti Utara Malaysia, Malaysia |
| Zakaria Kurdi, | University of Lynchburg, USA |
| Ze Tang, | Jiangnan University, China |
| Zewdie Mossie, | Debre Markos University, Ethiopia |
| Zhihan Lv, | Qingdao University, China |
| Zhou RouGang, | HangZhou DianZi University, China |
| Zoltan Gal, | University of Debrecen, Hungary |
| Zoran Bojkovic, | University of Belgrade, Serbia |

`

## Technically Sponsored by

**Computer Science & Information Technology Community (CSITC)**

**Artificial Intelligence Community (AIC)**

**Soft Computing Community (SCC)**

**Digital Signal & Image Processing Community (DSIPC)**

`

## 7th International Conference on Computer Science, Engineering and Information Technology (CSITY 2021)

## 7th International Conference on Networks & Communications (NWCOM 2021)

## 7th International Conference on Signal and Image Processing (SIGPRO 2021)

## 2nd International Conference on Advances in Software Engineering (ASOFT 2021)

## 7th International Conference on Artificial Intelligence and Fuzzy Logic Systems (AIFZ 2021)

`

## 2nd International Conference on Big Data & IOT (BDIoT 2021)

## 8th International Conference on Information Technology, Control, Chaos, Modeling and Applications (ITCCMA 2021)

## 2nd International Conference on Cloud Computing, Security and Blockchain (CLSB 2021)

## 7th International Conference on Data Mining (DTMN 2021)

`

## 2nd International Conference on Machine Learning Techniques and NLP (MLNLP 2021)

# INTRODUCING THE VIEWPOINT IN THE RESOURCE DESCRIPTION USING MACHINE LEARNING

Ouahiba Djama

Lire Laboratory, University of Abdelhamid
Mehri Constantine 2, Constantine, Algeria

### *ABSTRACT*

*Search engines allow providing the user with data and information according to their interests and specialty. Thus, it is necessary to exploit descriptions of the resources, which take into consideration viewpoints. Generally, the resource descriptions are available in RDF (e.g., DBPedia of Wikipedia content). However, these descriptions do not take into consideration viewpoints. In this paper, we propose a new approach, which allows converting a classic RDF resource description to a resource description that takes into consideration viewpoints. To detect viewpoints in the document, a machine learning technique will be exploited on an instanced ontology. This latter allows representing the viewpoint in a given domain. An experimental study shows that the conversion of the classic RDF resource description to a resource description that takes into consideration viewpoints, allows giving very relevant responses to the user's requests.*

### *KEYWORDS*

*Resource Description, RDF, Viewpoint, Ontology & Machine Learning.*

## 1. INTRODUCTION

With the rapid increase in the amount of published information and data on the web, search engines must select the most relevant information to the user's viewpoint. Thus, the resource description should take into consideration the different viewpoints of users. However, the existing descriptions on the web do not take into consideration the notion of the viewpoint. For that, we aim to convert the existing descriptions to descriptions that take into consideration the different viewpoints of users instead of building new descriptions.

Generally, on the web, the resource descriptions are available in RDF as DBPedia of Wikipedia content, etc. Djama [1] has proposed an RDF based framework (VP-RDF) to introduce the viewpoint in the description of resources.

Currently, no tool allows converting the existing RDF documents to VP-RDF documents. The conversion of the RDF document to VP-RDF document, allows introducing the viewpoint in the existing resource description.

For that, in this paper, we aim to propose an approach that allows converting RDF document to VP-RDF document.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 01-12, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111401

The RDF document belongs to a given domain. Thus, we will apply the machine learning technique on the instantiated multi-viewpoints ontology [2] of this domain to detect the relations between concepts / instances and viewpoints.

In the following section, we present a state of art. Then, in Section 3, we explain the proposed approach. After that, in Section 4 we apply the proposed approach to some use cases. Section 5 provides some results of the research. Finally, we present some areas for future work.

## 2. STATE OF THE ART

In this section, we present the definition of the viewpoint, some related work on the viewpoint and the machine learning with the ontology, VP-RDF language and the instantiated multi-viewpoints ontology.

### 2.1. Viewpoint Definitions

The authors have proposed different definitions of the notion of the viewpoint. In some works, the viewpoint corresponds to the perception of an object according to the observer's position [3]. For example, according to the observer's positions about the symbol '9', there are two viewpoints: number 6 and number 9 [2].

In some other work, (e.g., [1], [2], [4], [5] and [6]), the viewpoint is defined as a partial definition of an object basing on only some set of properties of this object. For example, in [2], the object apartment is defined by the size properties as area, room number, height, etc. and the finance properties as rent, price, etc. Therefore, we can give two different descriptions of the same apartment:

1) Viewpoint 1: large apartment, according to viewpoint 'size'.
2) Viewpoint 2: expensive apartment, according to viewpoint 'finance'.

Remark: Someone describes an apartment as cheap apartment and another one describes it as expensive apartment. This case is related to the fuzzy notion and not the viewpoint.

For example, the user's request aim to find all the existing properties of an apartment that describe its size [1]. With the exploitation of the classic descriptions, the search engine gives all the properties of this apartment [1] because it cannot detect that such properties are linked to the viewpoint size. This latter is because; the existing resource descriptions do not show the relations between the properties and the viewpoint [1].

The resource description that takes into consideration the viewpoint, allows linking each resource (property or entity) to a viewpoint [1].

### 2.2. Related Work

Several work are interested in the notion of the viewpoint. The authors in [5], [7], [8], [9], [10] and [11] have integrated the viewpoint in the development of the ontology. This ontology called 'multi-viewpoints ontology'.

Djezzar and Boufaida [12] have proposed an approach of the classification of an individual in the multi-viewpoints ontology. Djakhdjakha et al. [13] are interested in the alignment of the multi-viewpoints ontologies. Djama and Boufaida [2] have developed an approach that allows instancing the multi-viewpoints ontology.

These works are interested in the treatment of the ontology with viewpoints and not the treatment of the resources and documents with viewpoints.

Djama and Boufaida have also proposed an approach, in [14] and [15], which allows using multi-viewpoints ontology to annotate resources. This annotation allows describing the resource elements using the ontology elements. Then, the obtained annotation can be represented in RDF. Djama [1] has proposed the VP-RDF as an extension of the RDF to introduce the viewpoint in the description of resources.

Therefore, the works [1], [2], [14] and [15] allows constructing an RDF document from informal document as text document, XML document, etc. However, we aim to introduce the viewpoint in the existing RDF document. We will not construct an RDF document, but we will convert an existing RDF document to VP-RDF document.

Martin *et al.* [16] have exploited the viewpoints for reasoning on the classical ontology using case-based approach. In this work, the authors have not integrated the viewpoint in the ontology. Thus, the different viewpoints do not belong to a given domain. However, in our approach, the different viewpoints should appear in a given domain.

Gorshkov *et al.* [17] have exploited a multi-viewpoints ontology as a decision-making support. In this work, the authors have exploited the viewpoints that are represented in the ontology to make a decision. However, in our work, we aim to exploit the viewpoints that are represented in the instantiated multi-viewpoints ontology to recognize the set of the viewpoints in a given domain. These viewpoints will be introduced in the existing RDF documents. Thus, the resource description will be enriched.

Trichet *et al.* [18] have introduced the viewpoints in the semantic annotation of images. The authors have developed a platform that allows a user to use a set of ontologies to create a semantic annotation according to his/her viewpoint. The semantic annotation will be represented in RDF. However, the notion of the viewpoint cannot appear clearly in the RDF representation [1]. In the VP-RDF [1], the viewpoint appears clearly. Therefore, in our work, we aim to propose an approach that allows converting RDF document to VP-RDF document.

Several works allow introducing the context in RDF, as in [19], [20], [21] and [23]. These works allow representing that an assertion is true under a given context. For example [1], parallel lines can intersect in the context of solid geometry (3D geometry). However, in our work, we aim to represent the relation between a description of a resource and a viewpoint. For example, this is a large apartment; according to viewpoint ‘size’ and it is an expensive apartment, according to viewpoint ‘finance’.

According to Djama [1], the context and the viewpoint are two different notions. The context is a judgment based on rational arguments that represent a set of conditions [1] (Euclidean geometry or solid geometry). However, the viewpoint is a partial definition of an object [1].

Doan *et al.* [24] have developed a machine learning approach to establish semantic mappings among multiple ontologies. This approach is based on well-founded notions of semantic similarity, expressed in terms of the joint probability distribution of the concepts involved. The authors described the use of multi-strategy learning for computing concept similarities.

Kulmanov *et al.* [25] provided an overview over the methods that use ontologies to compute similarity and incorporate them in machine learning methods. The authors outline how semantic similarity measures and ontology embeddings can exploit the background knowledge in ontologies and how ontologies can provide constraints that improve machine-learning models.

The works [24] and [25] allow using machine-learning techniques to compute similarity between ontology concepts. However, we aim to make predictions of the relations between concepts/instances with the viewpoints.

## 2.3. VP-RDF

According to Djama [1], VP-RDF is an extension of the RDF by adding a new type of statement. The latter is composed of (Subject, Predicate_with_Viewpoint, Viewpoint) [1]. This statement allows linking a resource (Subject) to a viewpoint via the predicate Predicate_with_Viewpoint. To create this statement, Djama [1] proposed new elements that are shown in the table 1 and the table 2.

Table 1. VP-RDF Classes

| **Class** | **Definition by the RDF vocabulary** |
|---|---|
| VPrdf:Viewpoint | Subclass of “rdfs:Resource” |
| VPrdf:Predicate_with_Viewpoint | Subclass of “rdf:Property” |
| VPrdf:Statement | Subclass of “rdf:Statement” |

Table 2. VP-RDF Properties

| **Property** | **Domain** | **Range** |
|---|---|---|
| VPrdf:Subject_Statement | VPrdf:Statement | rdfs:Resources |
| VPrdf:Predicate_Statement | VPrdf:Statement | VPrdf:Predicate_with_Viewpoint |
| VPrdf:Object_Statement | VPrdf:Statement | VPrdf:Viewpoint |

## 2.4. Instantiated Multi-Viewpoints Ontology

The multi-viewpoints ontology [8] allows representing domain knowledge in two levels: Consensual and Heterogeneous level.

The consensual level allows representing the consensual concepts that are defined in all the viewpoints in the domain [2]. So, the consensual concepts can be considered not linked to particular viewpoints. These concepts are called global concept. Each of them is defined by a set of global properties (attributes) and a set of local properties [2]. A global property is defined in all the viewpoints [2]. So, a global property is consensual and can be considered not linked to particular viewpoints. Each local property is defined according to some viewpoints [2]. So, this property is linked to these viewpoints. The global concepts are organized in a global hierarchy [2].

For example [2], in the real estate domain, the global concept *Apartment* is defined by the properties: *surface, number of rooms, height,* etc. according to the viewpoint *Size*. The global concept *Apartment* is defined also by the properties: *price, taxes, rent price*, etc. according to the viewpoint *Finance.*

The value of a local property of a global concept allows generating a local concept according to the same viewpoint where this property is defined [2]. So, this concept is linked to this viewpoint. On the other hand, the global concept subsumes the local concept [2]. From this local concept, a hierarchy of local concepts will be built in the same viewpoint where this local concept is defined [2]. This local hierarchy of local concepts represents a local representation according to this viewpoint. The same global concept can be defined also by another local property that is defined in another viewpoint. In the same way, another local representation according of another viewpoint will be generated. And so on, the set of local representations represents the heterogeneous level.

For example [2], according to the property *surface*, we can generate the local concept *Large apartment*. This latter is defined according to the viewpoint *Size*. According to the property *price*, we can generate the local concept *Expensive apartment*. This latter is defined according to the viewpoint *Finance*. The global concept *Apartment* subsumes the local concept *Large apartment* and the local concept *Expensive apartment.*

The relationships between local concepts of the same viewpoint are called local roles [2]. Each relation is linked to the viewpoint where the local concepts are linked to. The relationships between local concepts of two different viewpoints are called global roles [2]. The global roles are not linked to particular viewpoints.

For example [2], the local role *rent* allows linking the local concepts *Rich tenant* and *expensive apartment* that are defined according to the viewpoint *Finance*. The local role *lives* allows linking the local concepts *Rich tenant* and *Large apartment*. *Rich tenant* is defined according to the viewpoint *Finance*. However, *Large apartment* is defined according to the viewpoint *size*.

The figure 1 shows the multi-viewpoints ontology structure.

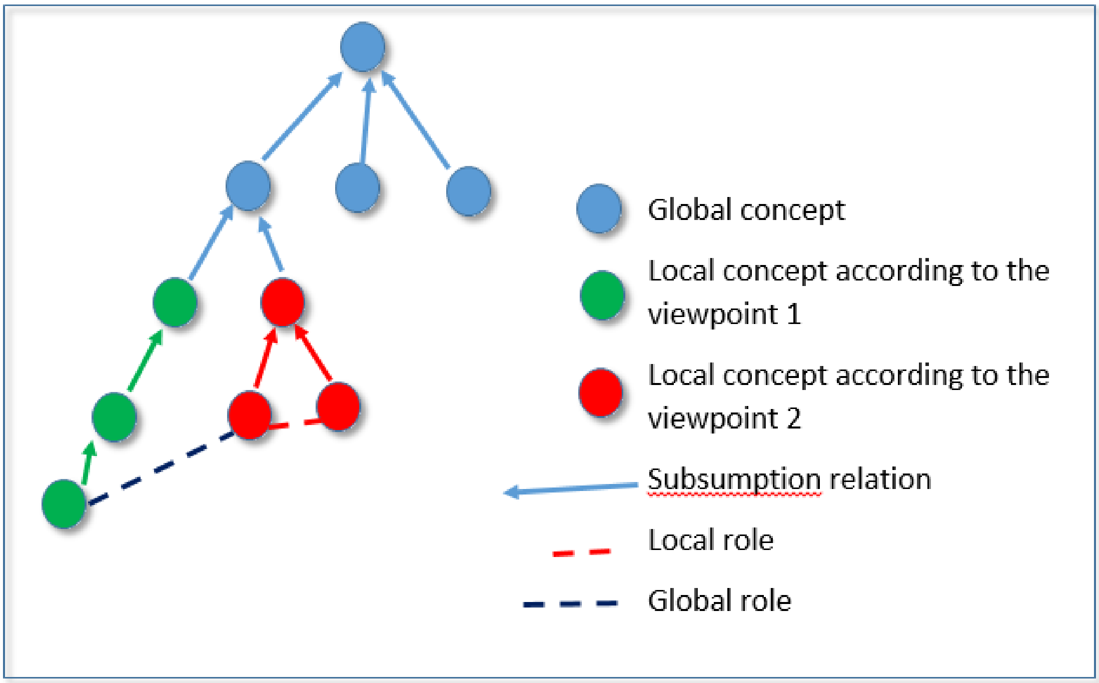

Global concept
Local concept according to the viewpoint 1
Local concept according to the viewpoint 2
Subsumption relation
Local role
Global role

Figure 1. The multi-viewpoints ontology structure

The instantiated multi-viewpoints ontology is a multi-viewpoints ontology that contains instances. An individual can be an instance of only one local concept under a particular viewpoint [2]. This individual can be an instance of another local concept in another viewpoint [2]. This individual is an instance of the global concept that subsumes all these local concepts.

The individual *apartment N°1* is an instance of the local concept *Large apartment* under the viewpoint *Size* and the local concept *Expensive apartment* under the viewpoint *Finance.* The individual *apartment N°1* is an instance of the global concept *Apartment.*

An individual, that is an instance of a local concept under a particular viewpoint, is linked to this viewpoint.

An individual, that is an instance of only a global concept, is not linked to particular viewpoints. Therefore, the instantiated multi-viewpoints ontology allows representing knowledge in two levels. The consensual knowledge, which are not linked to viewpoints in the domain, are represented in the consensual level. The knowledge, which are linked to viewpoints, are represented in the heterogeneous level. We resume the descriptions of the components of the instantiated multi-viewpoints ontology in the table 3.

Table 3. Components of the instantiated multi-viewpoints ontology

| **Element** | **Description** |
|---|---|
| Global concept | Not linked to viewpoints |
| Global attribute (data type property) | Not linked to viewpoints |
| Global role (object property) | Not linked to viewpoints |
| Global instance | Not linked to viewpoints |
| Local concept | Linked to one or several viewpoints |
| Local attribute (data type property) | Linked to one or several viewpoints |
| Local role (object property) | Linked to one or several viewpoints |
| Local instance | Linked to one or several viewpoints |

## 3. PROPOSED APPROACH

The proposed approach aims to convert an RDF document to VP-RDF document. For that, it is composed of two main steps:

### 3.1. Detection of Viewpoints

An RDF document is composed of a set of RDF triplets (statements). Each statement is composed of subject, predicate and object. This step aims to detect for each statement and for each its component the viewpoint that is linked to. There are four main cases:

1) Subject of the statement is linked to one or several viewpoints.
2) Object of the statement is linked to one or several viewpoints.
3) Both the subject and the object are linked to one or several viewpoints.
4) Predicate of the statement is linked to one or several viewpoints.

The other subcases are the combination between these cases.

In this step, we aim to make predictions:

a) which viewpoint that the RDF subject is linked to,
b) which viewpoint that the RDF object is linked to,
c) which viewpoint that the RDF predicate is linked to.

To make predictions, it is necessary to exploit a machine learning techniques. The RDF document belongs to a given domain. Thus, we will exploit the instantiated multi-viewpoints ontologies [2] of this domain to extract viewpoints in the domain. Therefore, we create a model that will learn, from various instantiated multi-viewpoints ontologies of the same domains, to link a term to a viewpoint. A term can be a concept or an instance of a concept. This model will learn also to link a relation between terms to a viewpoint. Therefore, we can exploit a machine learning technique that is based on the computing of the frequencies of relations between terms and viewpoints.

After the machine-learning step, the created model is able to predict (detect) the viewpoint that the RDF subject / the RDF object / the RDF predicate is linked to. We choose to use a machine learning technique to detect a viewpoint in order to avoid reasoning on the instantiated multi-viewpoints ontologies, each time we want to convert an RDF document to a VP-RDF document. The reasoning on the ontologies takes a considerable amount of time.

Now, the created model will be exploited in the first step in our approach to detect:

1) The viewpoint/ the set of viewpoints where the subject of the statement is linked to.
2) The viewpoint/ the set of viewpoints where the object of the statement is linked to.
3) The viewpoint/ the set of viewpoints where the predicate of the statement is linked to.

For example, in the real estate domain, *Rich_Tenant* is always linked to the viewpoint *Finance. Large_Apartment* is linked to the viewpoint *size.* The relation *rent* is linked to the viewpoint *Finance.*

For example, in the education domain, *Professor* is linked to the viewpoint *University_Education.*

## 3.2. Creation of VP-RDF Statements

This step aims to exploit the VP-RDF vocabulary and syntax [1] (see subsection 2.3.) to convert the RDF statement to VP-RDF statement. There are four cases:

1. Subject of the statement is linked to one or several viewpoints: in this case, the RDF statement will be kept and from the subject, one or several VP-RDF statements will be created. For example, the RDF statement (<Rich_Tenant>,<lives_in>,<Constantine>) will be converted in VP-RDF in two statements:
   (<Rich_Tenant>,<lives_in>,<Constantine>) and
   (<Rich_Tenant>,<belong_to>,<Finance>)).
   From the subject 'Rich_Tenant', one VP-RDF statement will be created; because 'Rich_Tenant' is linked to only one viewpoint.
2. Object of the statement is linked to one or several viewpoints: in this case, the RDF statement will be kept and from the object, one or several VP-RDF statements will be created.
3. Both the subject and the object are linked to one or several viewpoints: in this case, the RDF statement will be kept. From the subject, one or several VP-RDF statements will be created. From the object, one or several VP-RDF statements will be created.

4. Predicate of the statement is linked to one or several viewpoints: in this case, first, it is necessary to create a class to represent the predicate. Then, this class will be linked to a viewpoint via a VP-RDF statement. It will be linked also to the object of the RDF statement via a new RDF statement, where this class becomes a subject of the new RDF statement.

The figure 2 shows the creation of the machine-learning model. The figure 3 shows the different steps of the proposed approach with the inputs and outputs.

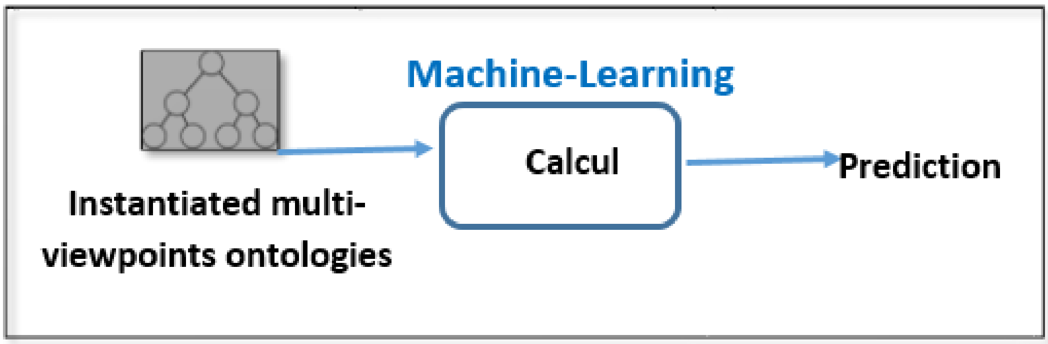


Figure 2. The creation of the machine-learning model

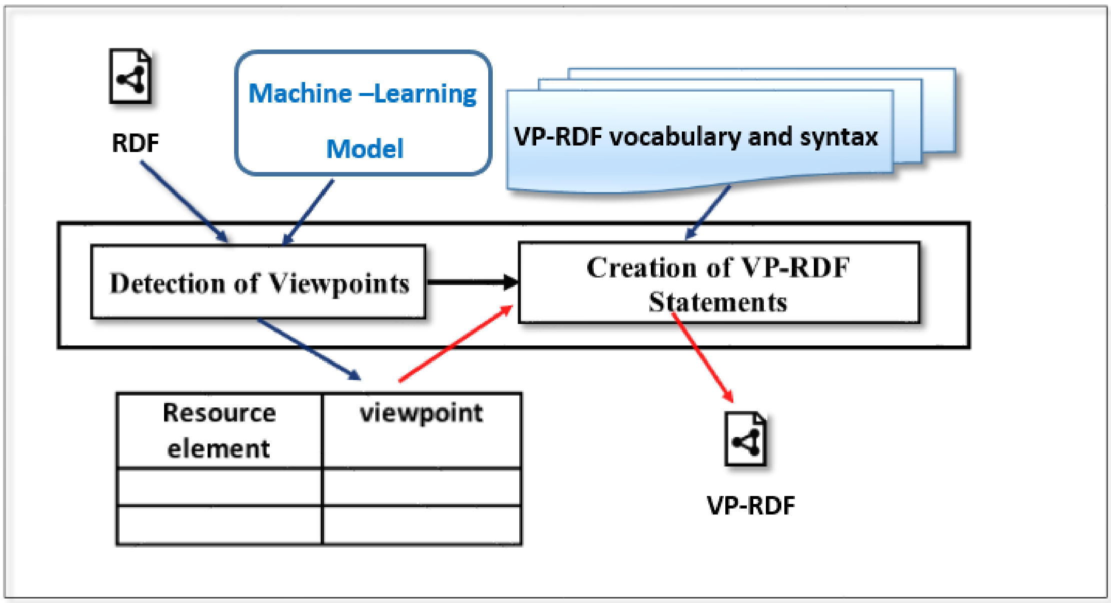


Figure 3. Schema of the proposed approach

## 4. Use Cases

In this section, we present an example for each case seen in the previous section. These examples are regrouped in the following table:

Table 4. Use Cases

| **Statement in RDF Document** | **Conversion to VP-RDF** |
|---|---|
| (<Rich_Tenant>,<lives_in>,<Constantine>) | (<Rich_Tenant>, <lives_in>, <Constantine>)<br>(<Rich_Tenant >, < belong_to>, <Finance>) |
| (<John>,<is>, <Professor >) | (<John>, <is>, <Professor >)<br>(<Professor > <defined_according _to>, <University_Education >) |
| (<Rich_Tenant>,<lives_in>,<Large_Apartment>) | (<Rich_Tenant>, <lives_in>, <Large_Apartment>)<br>(<Rich_Tenant >, < belong_to>, <Finance>)<br>(<Large_Apartment> <according_to> <size>) |
| (<John>,<rent>, <apartment3>) | (<John>,<rent>, <apartment3>)<br>(<Class_rent >, < rent _value>, < apartment3>)<br>(<Class_rent >, < belong_to>, <Finance>) |

In this table, the statements that are written with red colour represent VP-RDF statements in the VP-RDF language. The black ones represent RDF statements in the VP-RDF language.

## 5. Results and Discussion

We have realized an experimental study to evaluate the effect of the exploitation of the VP-RDF documents in the search of responses to the user's requests. We have converted some pre-existed RDF documents to VP-RDF documents. The RDF documents belong to different domains: 150 RDF documents of the real estate domain, 125 RDF documents of the library domain, 175 RDF documents of the education domain and 1200 RDF documents of the medical domain.

We asked 200 users to make a request for each these domains according to their interests (viewpoints). We exploited the VP-RDF documents to give responses to the user's requests. We asked the same users to evaluate these responses if are relevant to their interests. Finally, we calculated the percentage of the number of responses that are relevant to the user's interest. The results are regrouped into the table 5:

Table 5. Results of the study

| **Domain** | **Percentage of the Relevant Responses** |
|---|---|
| Real estate | 98.1% |
| Library | 98.3% |
| Education | 98% |
| Medical | 94.4% |

We find that all the percentages are high in all the domains. This is because the relationships between viewpoints and resources are explicitly represented in the VP-RDF documents.

For example, in the real estate domain, a user wants to know all the existing properties of an apartment that describe its size [1].

In VP-RDF, the properties (e.g., height, *surface*, the number of rooms, etc.) are linked directly (explicitly) to the viewpoint size [1]. The answer shows the values of height, *surface*, the number of rooms, etc. So, the answer will be found directly without reasoning. However, in the RDF, the properties (e.g., height, *surface*, the number of rooms, etc.) are not linked to the viewpoint size. The answer shows the values of the all properties of the apartment. In order to improve the answer, it is necessary to exploit a reasoning tool to select only the values of the properties: height, *surface*, the number of rooms, etc. However, the reasoning can give the results that are not accurate. For example, the answer can be the value of the properties: *height* and *surface*. However, *number of rooms* will not be mentioned. Therefore, with classic RDF, the answer can be incomplete or more than the user's needs. In the last case, the user should select the relevant answer manually.

With VP-RDF, the answer will be completely relevant exactly to the user's needs. Therefore, the user does not need to select the relevant answer manually.

## 6. Conclusions

We have proposed an approach that allows converting RDF document to VP-RDF document. This latter allows introducing explicitly the viewpoint in the description of the resources and their relationships [1]. This description helps search engines to give responses that are relevant to the user's interest (viewpoint) with high rate and the user does not need to select the relevant answer manually. The machine learning mechanism is exploited to detect viewpoints basing on the instantiated multi-viewpoints ontology.

The accuracy of the proposed method depends on the accuracy of the detection of the viewpoints in a given domain. It depends also on the accuracy of the predictions about the relation between resources and viewpoints. Therefore, the amount and the quality of the data exploited in the machine learning affect the accuracy of the proposed approach.

The imprecision and the uncertainty can be coupled with the viewpoints in the resource description. For example, in the finance viewpoint, someone describes an apartment as cheap apartment and another one describes it as expensive apartment. This case, the two persons share the same viewpoint (Finance). In addition to this, the descriptions are related to the fuzzy notion too. As future work, we plan to couple the viewpoint with the fuzzy in the description of resources.

### References

[1] O. Djama, (2021) "VP-RDF: an RDF based framework to introduce the viewpoint in the description of resources," *Applied Computer Systems*, Vol. 26, No. 1, pp. 44–53. https://doi.org/10.2478/acss-2021-0006

[2] O. Djama & Z. Boufaida, (2020) "Instantiation of the multi-viewpoints ontology from a resource," *International Journal of Computers and Applications*, pp. 1–12. https://doi.org/10.1080/1206212X.2020.1711615

[3] M. Minsky, (1975) "A framework for representing knowledge," *in The Psychology of Computer Vision*, P. H. Winston, Ed. New York : McGraw-Hill, pp. 211–217.

[4] O. Marino, (1993) “Raisonnement classificatoire dans une représentation à objets multi-points de vue,” PhD thesis, University of Joseph Fourier, Grenoble, France.

[5] L. T. Bach, (2006) “Construction d'un web sémantique multi-points de vue,” Ph.D. thesis, Mines ParisTech (école des Mines de Paris), Sophia Antipolis, France.

[6] G. Falquet & C. L. Mottaz, (2001) “Navigation hypertexte dans une ontologie multi-points de vue,” presented at the conference NîmesTIC'2001, Nîmes, France.

[7] M. Zhitomirsky-Geffet, E. S. Erez & B.I. Judit (2017) “Toward multiviewpoint ontology construction by collaboration of nonéexperts and crowdsourcing: the case of the effect of diet on health,” *Journal of the Association for Information Science and Technology*, Vol. 68, No. 3, pp. 681–694. https://doi.org/10.1002/asi.23686

[8] M. Hemam, (2018) “An extension of the ontology web language with multi-viewpoints and probabilistic reasoning,” *International Journal of Advanced Intelligence Paradigms*, Vol. 10, No. 3, pp. 247–265. https://doi.org/10.1504/IJAIP.2018.090789

[9] M. Hemam, M. Djezzar & Z. Boufaida, (2017) “Multi-viewpoints ontological representation of composite concepts: A description logics-based approach,” *International Journal of Intelligent Information and Database Systems*, Vol. 10, No.1/2, pp. 51–68. https://doi.org/10.1504/IJIIDS.2017.086193

[10] J. Kingston, (2008) “Multi-Perspective Ontologies: Resolving Common Ontology Development Problems,” *Expert Systems with Applications*, Vol. 34, No. 1, pp. 541–550. https://doi.org/10.1016/j.eswa.2006.09.040

[11] E. Zemmouri, H. Behja, A. Marzak & B. Trousse, (2014) “Ontology-Based Knowledge Model for Multi-View KDD Process,” *International Journal of Mobile Computing and Multimedia Communications*, Vol. 4, No. 3, pp. 21–33. https://doi.org/10.4018/jmcmc.2012070102

[12] M. Djezzar & Z. Boufaida, (2015) “Ontological classification of individuals: a multi-viewpoints approach,” *International Journal of Reasoning-based Intelligent Systems*, Vol. 7, No. 3/4, pp. 276–285. https://doi.org/10.1504/IJRIS.2015.072954

[13] L. Djakhdjakha, M. Hemam & Z. Boufaida, (2014) “Towards a representation for multi-viewpoints ontology alignments,” *International Journal of Metadata, Semantics and Ontologies*, Vol. 9, No. 2, pp. 91–102. https://doi.org/10.1504/IJMSO.2014.060324

[14] O. Djama & Z. Boufaida, (2013) “Multi-viewpoints semantic annotation of XML documents,” *in Proceedings of the World Congress on Engineering 2013, International Association of Engineers: 2013*, S. I. Ao, L. Gelman, D. W. L. Hukins, A. Hunter and A. M. Korsunsky, Eds. London (U.K): Newswood Limited, pp. 390–394.

[15] O. Djama (2020) “Annotation sémantique Multi-Points de Vue (MPV) de ressources et leur exploitation à travers un langage basé RDF,” Ph.D. thesis, University of Abdelhamid Mehri-Constantine 2, Constantine, Algeria.

[16] A. Martin, S. Emmenegger, K. Hinkelmann & B. Thonssen, (2017) “A Viewpoint-Based Case-Based Reasoning Approach Utilising an Enterprise Architecture Ontology for Experience Management,” *Enterprise Information Systems*, Vol. 11, No. 4, pp. 551–575. https://doi.org/10.1080/17517575.2016.1161239

[17] S. Gorshkov, S. Kralin & M. Miroshnichenko, (2016) “Multi-Viewpoint Ontologies for Decision-Making Support*,” in Knowledge Engineering and Semantic Web. KESW 2016. Communications in Computer and Information Science*, A. C. Ngonga Ngomo and P. Kremen, Eds. Suisse (Switzerland): Springer-Cham, pp. 3–17. https://doi.org/10.1007/978-3-319-45880-9_1

[18] F. Trichet, X. Aimé & C. Thovex, (2010) “OSIRIS: Ontology-Based System for Semantic Information Retrieval and Indexation Dedicated to Community and Open Web Spaces,” *in Hershey Handbook of Research in Culturally-Aware Information Technology: Perspectives and Models*, E. G. Blanchard and D. Allard, Eds. PA: Information Science Publishing, pp. 465–483. https://doi.org/10.4018/978-1-61520-883-8.ch021

[19] A. Analyti, C. V. Damasio & I. Pachoulakis, (2015) “Nested Contextualised Views in the Web of Data,” *International Journal of Web Engineering and Technology*, Vol. 10, No. 1, pp. 31–64. https://doi.org/10.1504/IJWET.2015.069360

[20] H. Cherfi, O. Corby, C. Faron-Zucker, K. Khelif & M. T.Nguyen, (2008) “Semantic Annotation of Texts with RDF Graph Contexts,” presented at Conceptual Structures: Knowledge Visualization and Reasoning. International Conference on Conceptual Structures (ICCS’08), CEUR-WS, Toulouse, France.

[21] O. Khriyenko & V. Terziyan, (2006) "A Framework for Context-Sensitive Metadata Description," *International Journal of Metadata, Semantics and Ontologies*, Vol. 1, No. 2, pp. 154 –164. https://doi.org/10.1504/IJMSO.2006.011011

[22] H. Stoermer, P. Bouquet, I. Palmisano & D. Redavid, (2007) "A Context-Based Architecture for RDF Knowledge Bases: Approach, Implementation and Preliminary Results," *in Web Reasoning and Rule Systems*, M. Marchiori, J. Z. Pan and C. S. Marie, Eds. Berlin (Heidelberg): Springer, pp. 209–218. https://doi.org/10.1007/978-3-540-72982-2_15

[23] R. Vanlande & C. Nicolle, (2008) "Context DataModel Framework: Semantic Facilities Management," *International Journal of Product Lifecycle Management*, Vol. 3, No. 2/3, pp. 165–177. https://doi.org/10.1504/IJPLM.2008.021440

[24] A. Doan, J. Madhavan, P. Domingos & A. Halevy, (2004) "Ontology Matching: A Machine Learning Approach", In *Handbook on Ontologies,* S. Staab, and R. Studer, Eds. Berlin: Springer, pp 385-403.

[25] M. Kulmanov, F. Z. Smaili, X. Gao & R. Hoehndorf (2020) "Semantic similarity and machine learning with ontologies", *Briefings in Bioinformatics*, pp. 1–18. https://doi.org/10.1093/bib/bbaa199

## AUTHOR

**Ouahiba Djama** was born in Constantine, Algeria. She received her Engineer degree in Computer Science in 2005 and M. sc. in 2010, both from University of Mentouri, Constantine, Algeria. She obtained a Doctoral degree in Computer Science in 2020 from the University of Abdelhamid Mehri Constantine 2, Ali Mendjeli, Constantine, Algeria. She is currently an Assistant Professor at the University of Mentouri Brothers-Constantine 1, Constantine, Algeria. She is also an Attached Member of the SI&BC research group at Lire Laboratory of the University of Abdelhamid Mehri-Constantine 2, Constantine, Algeria. Her research interests include knowledge representation and reasoning, formal knowledge representation for semantic web, ontology development, web technologies, big data, bioinformatics, artificial intelligence and computer applications.
djama_ouahiba@umc.edu.dz

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# SMARTBARK: AN AUTOMATED DOG BARKING DETECTION AND MONITORING SYSTEM USING AI AND DEEP LEARNING

Xiangjian Liu[1], Yishan Zou[2] and Yu Sun[3]

[1]Arnold O. Beckman High School
[2]University of Pennsylvania, Philadelphia, PA 19104
[3]California State Polytechnic University, Pomona, CA 91768

## *ABSTRACT*

*Dogs have the tendency to bark at loud noises that they perceive as an intruder or a threat, and the hostile barking can often last up to hours depending on the duration of such noise. These barking sessions are unnecessary and negatively impact the quality of life of the others in your community, causing annoyance to your neighbors [1]. Having the rights to file noise complaints to the Home Owners Association, potentially resulting in fines or even the removal of the pet [2]. In this paper, we will discuss the development of an algorithm that takes in audio inputs through a microphone, then processes the audio and identifies that the audio clip is dog barks through machine learning, and ultimately sends the notification to the user. By implementing our application to the everyday life of dog owners, it allows them to accurately determine the status of their dog in real-time with minimal false reports.*

## *KEYWORDS*

*Monitoring System, Deep Learning, AI, mobile platform*

## 1. INTRODUCTION

Noise pollution, or excessive unwanted sound, is a problem worldwide, causing nuisances ranging from annoyance, mental breakdowns, to deafness [3]. In 1981, the Environmental Protection Agency estimated that nearly 50% of the people in the United States were exposed to harmful noise pollution [5]. Consequently, homeowners would contact local authorities to remove sources of noise pollution [2]. Without proper communication between homeowners, these disputes could often end up in unnecessary fines and penalizations [4] in preventable situations, especially against dog owners. Dogs, as tame, loyal pets, could be easily controlled by their owner's commands. Their barks, however, are purely instinctive and often unable to be trained.

Some of the pre-existing solutions have been proposed regarding sound event detection (SED), most of which allow the user to check in on their home environment in real-time. More advanced monitors offer the functionality of recording sound files and storing them in a database, allowing the user to access archived sound files. These proposals offer a decibel based detection method, meaning that sound will be recorded if the loudness of the sound is over a certain decibel. However, these solutions require the user to determine the sound and are incapable of identifying sound files alone.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 13-25, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111402

Our method follows the same premise in monitoring noise in real-time along with some additional features that are crucial to the user experience. First, our method allows for accurate and precise classification [6] of sound files of 10 categories, specifically dog barks. Though our method is exclusive for dog barks, it provides a high customizability so that it could be specialized for classification and notification of other sounds. Second, the implementation of cloud messaging [7] and app notifications lets the user know if the barking is ensuing in real-time. Each entry of identified dog bark is recorded with corresponding confidence score [8] and time recorded, and transmitted to the user's application, allowing the user to identify the need to go back.

With the goal of achieving real-time prediction and notification in mind, we also needed to minimize false predictions by producing an efficient model. In two application scenarios, we demonstrate how the variation of elements during model training substantially impacts the program's performance during testing. First, we prove the effectiveness of utilizing a larger set of training data by comparing the accuracy of models trained by datasets of different sizes. Second, we weigh the impact of including additional categories of prediction results on prediction accuracy. Through trials of experiment, we compare the mean prediction accuracy of each variation in these two experiments in order to find the most efficient arrangement of these elements.

The rest of the paper is organized as follows: Section 2 scrutinizes the challenges that we faced during the experiment and design process; Section 3 provides detail on our solutions to the corresponding challenges mentioned in Section 2; Section 4 presents relevant details regarding the experiment, followed by presenting related work in Section 5. Finally, Section 6 concludes this research paper, as well as outlining future work of this project.

## 2. CHALLENGES

### 2.1. Challenge 1: How to Build the Training Model Without Having the Available Data From the Specific Dogs

In order to build a model with high accuracy and precision, a vast amount of data would be needed while training the model. This poses a challenge to our team as not only do we have to record hundreds to thousands of sound files of different categories, these sound files would not provide enough diversity for the program to work accurately on the dog barks of different breeds

### 2.2. Challenge 2: How to Perform Real-Time Bark Sound Detection

With the premise of notifying users the current state of their dogs, our project must perform both data collection and prediction in real-time. With each sound file recorded, a corresponding prediction needs to be made without delay to ensure the resulting predictions reflect the condition of the environment at all times. At first glance, the usage of a cloud server becomes a possibility as it provides ample processing power to ensure low prediction run time, but local predictions also provide accurate predictions without server delay.

### 2.3. Challenge 3: How to Integrate Components From Different Platforms With a Low Latency

Data collection, sound processing, prediction, and notification constitute this system, and the integration of these components with low latency poses another challenge for us. Aside from creating a system with real-time bark detection and prediction, the user should also be notified in real-time when their dogs are barking. To accomplish this goal, the usage of a database is necessary for the retrieval and storage of data, and the user's mobile device should be connected to the database for easy access. However, the efficiency of these methods are yet to be determined for this system to be declared as real-time.

# 3. SOLUTION

## 3.1. Overview of the Solution

In order to perform predictions on audio inputs and notify the user in real-time, a three-part system was developed. As shown in Figure 1, The Dog Bark Detector is a system that records audio inputs in 5-second intervals through a microphone on a Raspberry Pi, which then performs predictions locally through a trained model on every sound file. If the program predicts that the sound file signifies a dog bark, information such as the time of the event and the confidence score is then transmitted to the online database, as the Raspberry Pi is connected to the internet. Along with the storage of that information in the database, a notification is also sent to the user's mobile device, notifying them that their dog is likely barking at the moment. The user is also able to access all of the archived information in the database on their mobile application.

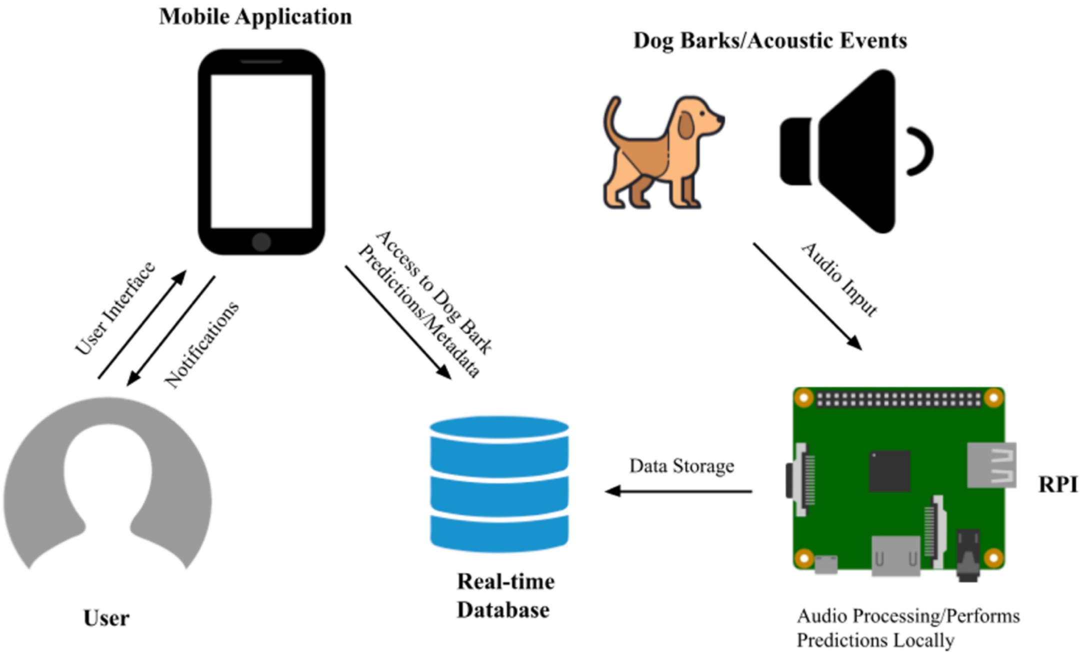

Mobile Application
Dog Barks/Acoustic Events
User Interface
Notifications
Access to Dog Bark Predictions/Metadata
Audio Input
Data Storage
RPI
User
Real-time Database
Audio Processing/Performs Predictions Locally

Figure 1. Graphic Representation of the solution

## 3.2. Machine Learning

### a. Introduction

To get started with the core component of this system, we needed to find an available dataset to train our machine with, as recording thousands of sound files and manually labeling them is a time-consuming process. Furthermore, we needed to find an adequate library for predictions.

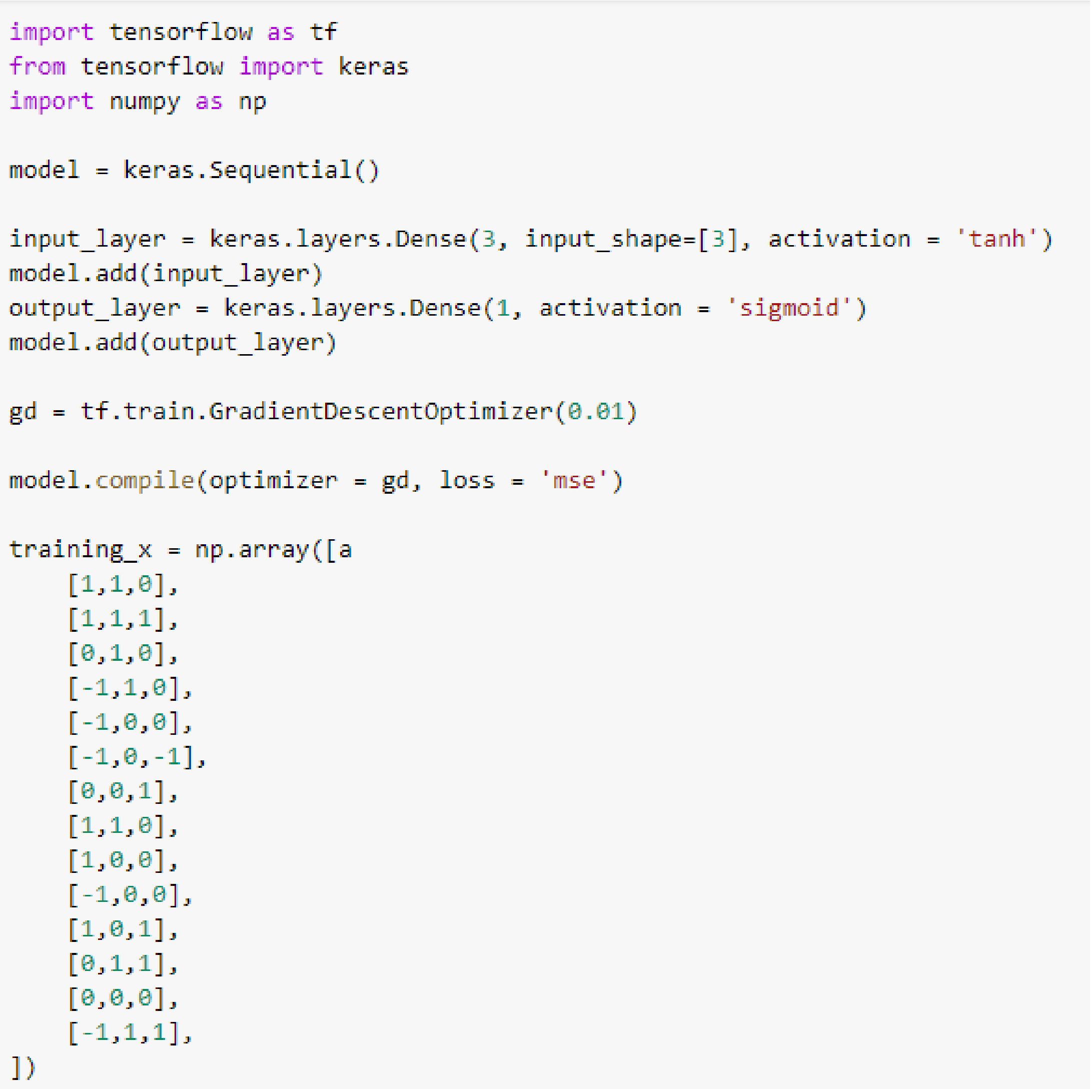
## b. Dataset

In order to have enough training data to produce a model with both accuracy and precision, a copious amount of diverse, categorized sound files would be needed. However, achieving this goal would not be possible in a reasonable amount of time, so we utilized an online dataset. UrbanSound8K [9] provides a variety of training data of 10 categories, up to 8732 sound files.

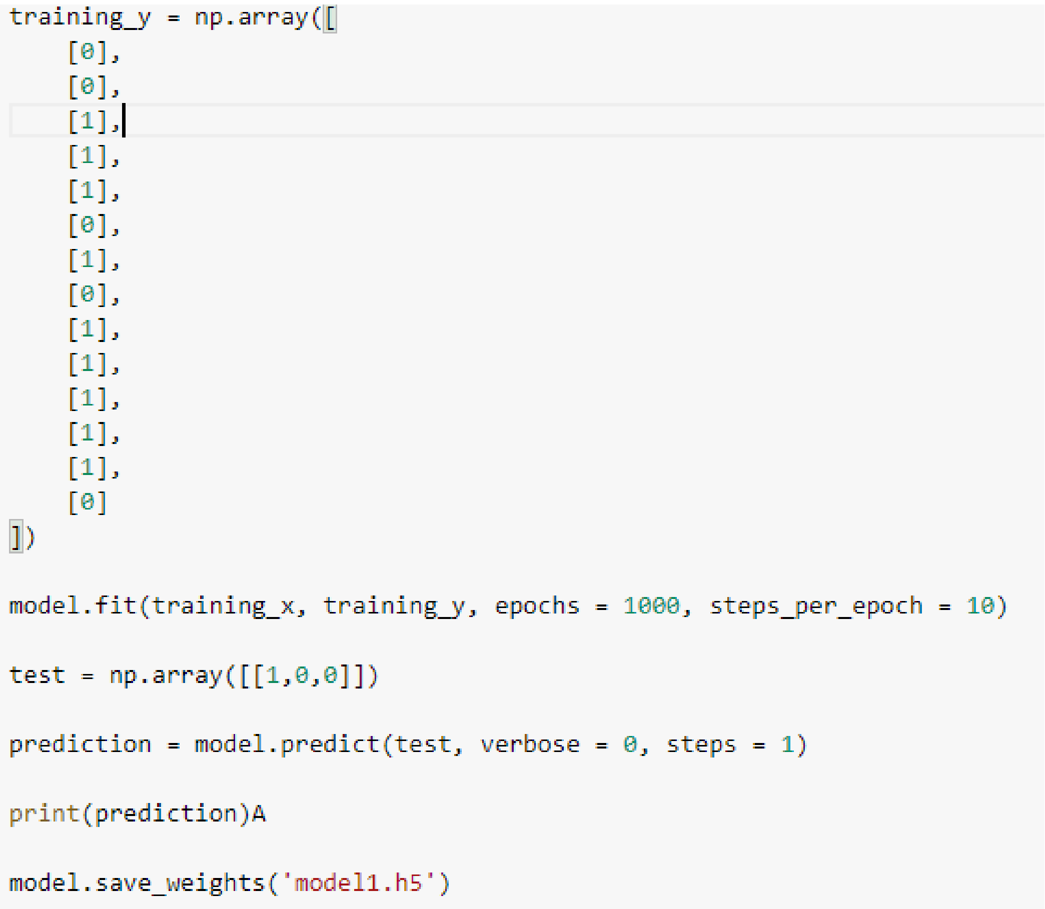
## c. Training Model

## d. Attach Some Code

The following Figure 2 shows the utilization of the Neural Network during model training.



```
import tensorflow as tf
from tensorflow import keras
import numpy as np

model = keras.Sequential()

input_layer = keras.layers.Dense(3, input_shape=[3], activation = 'tanh')
model.add(input_layer)
output_layer = keras.layers.Dense(1, activation = 'sigmoid')
model.add(output_layer)

gd = tf.train.GradientDescentOptimizer(0.01)

model.compile(optimizer = gd, loss = 'mse')

training_x = np.array([a
    [1,1,0],
    [1,1,1],
    [0,1,0],
    [-1,1,0],
    [-1,0,0],
    [-1,0,-1],
    [0,0,1],
    [1,1,0],
    [1,0,0],
    [-1,0,0],
    [1,0,1],
    [0,1,1],
    [0,0,0],
    [-1,1,1],
])
```



```
training_y = np.array([
    [0],
    [0],
    [1],
    [1],
    [1],
    [0],
    [1],
    [0],
    [1],
    [1],
    [1],
    [1],
    [1],
    [0]
])

model.fit(training_x, training_y, epochs = 1000, steps_per_epoch = 10)

test = np.array([[1,0,0]])

prediction = model.predict(test, verbose = 0, steps = 1)

print(prediction)A

model.save_weights('model1.h5')
```

Figure 2. Model Training and Neural Net

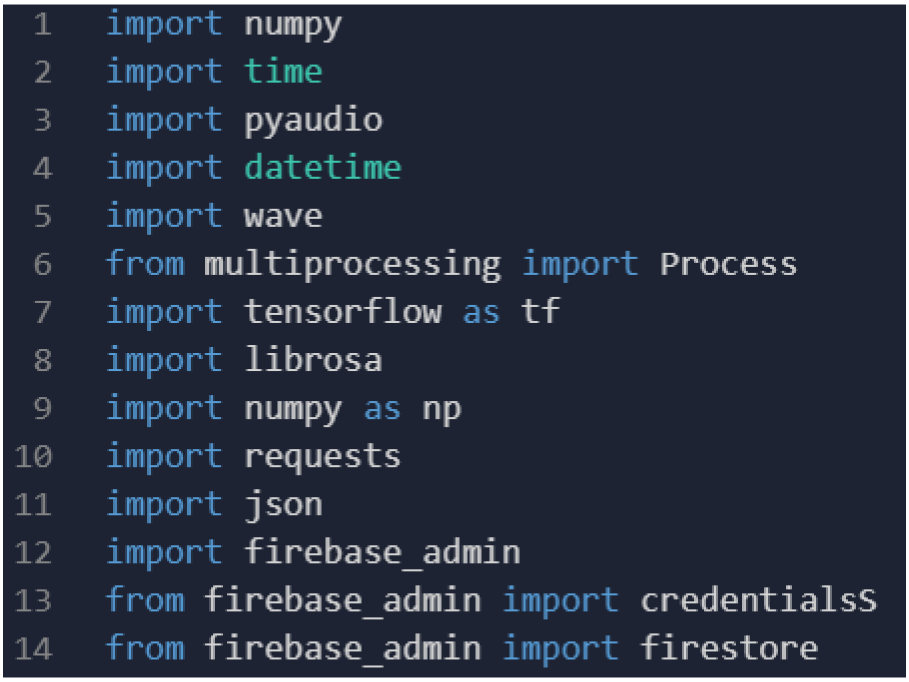
## 3.3. Raspberry PI

### a. Installation

During the installation process of the Raspberry Pi, we decided to include the addition of an extended USB adapter cord. The reasoning behind this is that the Raspberry Pi 4 includes a cooling fan to prevent overheating over a prolonged period of runtime. If the microphone for audio input were to be connected to the Raspberry Pi IO directly, the fan-produced background noise interferes with the audio, and could potentially throw off the predictions.

### b. Libraries

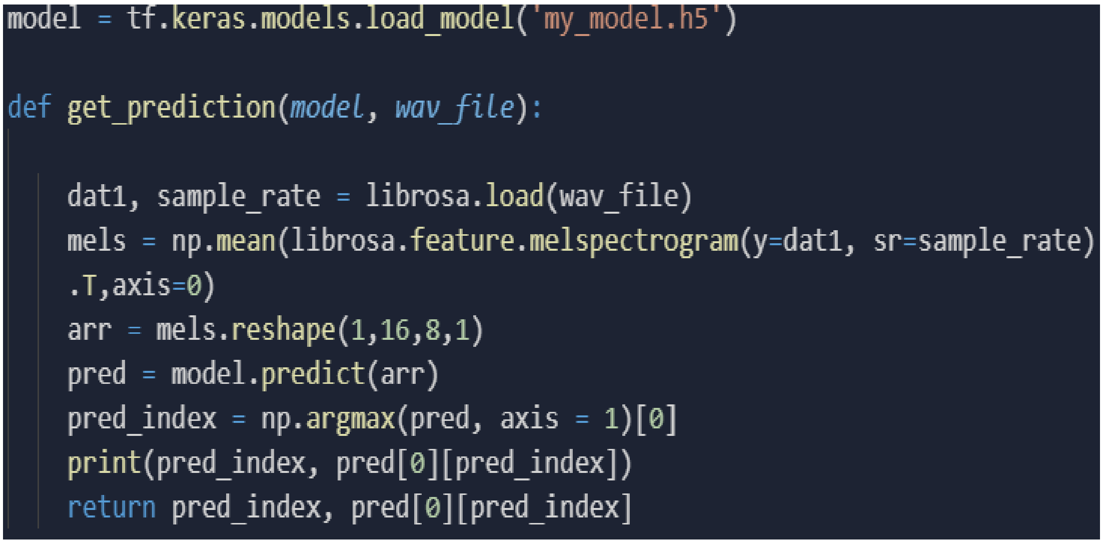
Figure 3 shows the library integrated into the program, while Figures 4 and 5 present snippets of the program.



```python
import numpy
import time
import pyaudio
import datetime
import wave
from multiprocessing import Process
import tensorflow as tf
import librosa
import numpy as np
import requests
import json
import firebase_admin
from firebase_admin import credentialsS
from firebase_admin import firestore
```

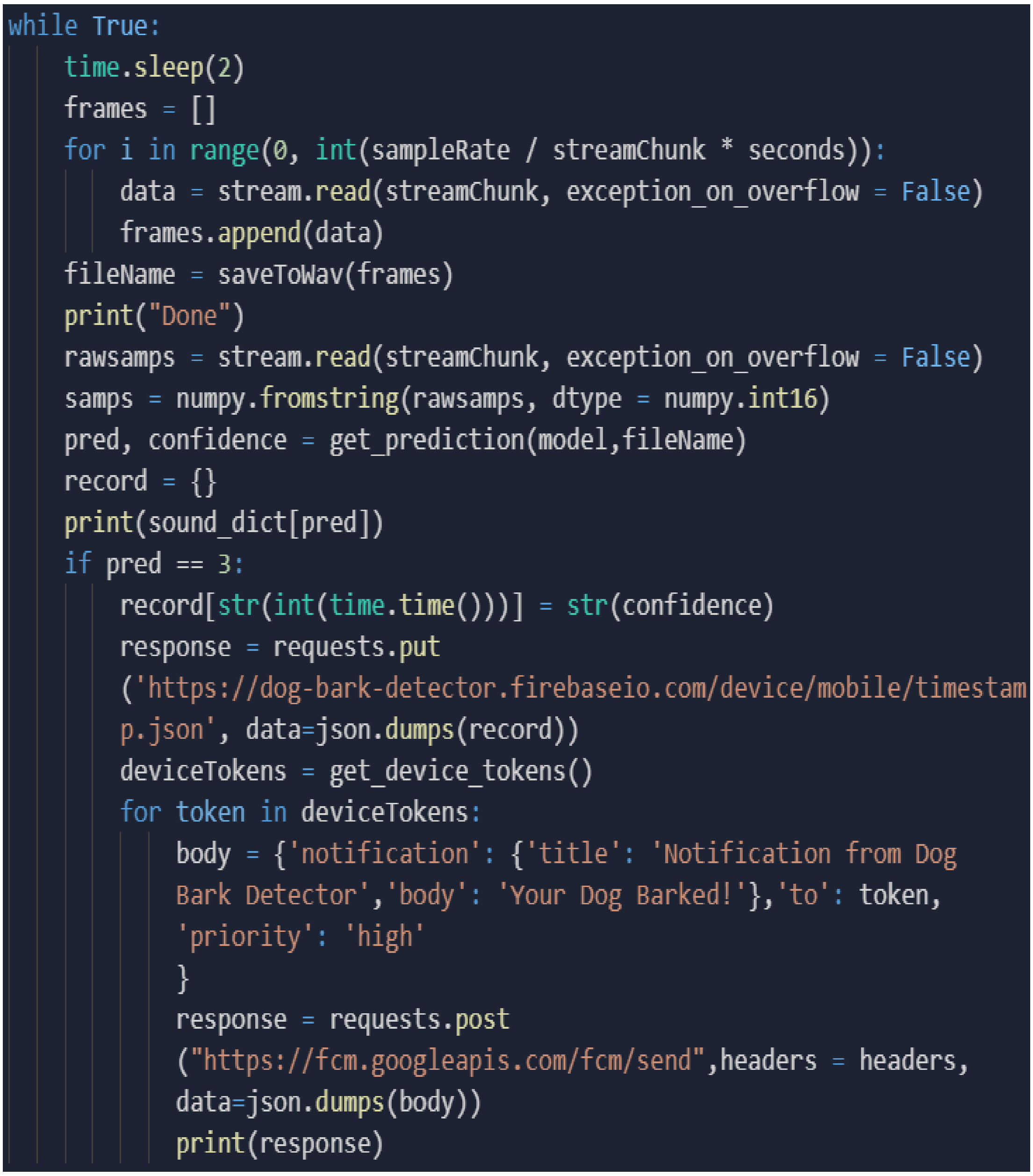
Figure 3. Imported Libraries

### c. Code



```python
model = tf.keras.models.load_model('my_model.h5')

def get_prediction(model, wav_file):

    dat1, sample_rate = librosa.load(wav_file)
    mels = np.mean(librosa.feature.melspectrogram(y=dat1, sr=sample_rate)
    .T,axis=0)
    arr = mels.reshape(1,16,8,1)
    pred = model.predict(arr)
    pred_index = np.argmax(pred, axis = 1)[0]
    print(pred_index, pred[0][pred_index])
    return pred_index, pred[0][pred_index]
```

Figure 4. Prediction Function



```
while True:
    time.sleep(2)
    frames = []
    for i in range(0, int(sampleRate / streamChunk * seconds)):
        data = stream.read(streamChunk, exception_on_overflow = False)
        frames.append(data)
    fileName = saveToWav(frames)
    print("Done")
    rawsamps = stream.read(streamChunk, exception_on_overflow = False)
    samps = numpy.fromstring(rawsamps, dtype = numpy.int16)
    pred, confidence = get_prediction(model,fileName)
    record = {}
    print(sound_dict[pred])
    if pred == 3:
        record[str(int(time.time()))] = str(confidence)
        response = requests.put
        ('https://dog-bark-detector.firebaseio.com/device/mobile/timestam
        p.json', data=json.dumps(record))
        deviceTokens = get_device_tokens()
        for token in deviceTokens:
            body = {'notification': {'title': 'Notification from Dog
            Bark Detector','body': 'Your Dog Barked!'},'to': token,
            'priority': 'high'
            }
            response = requests.post
            ("https://fcm.googleapis.com/fcm/send",headers = headers,
            data=json.dumps(body))
            print(response)
```

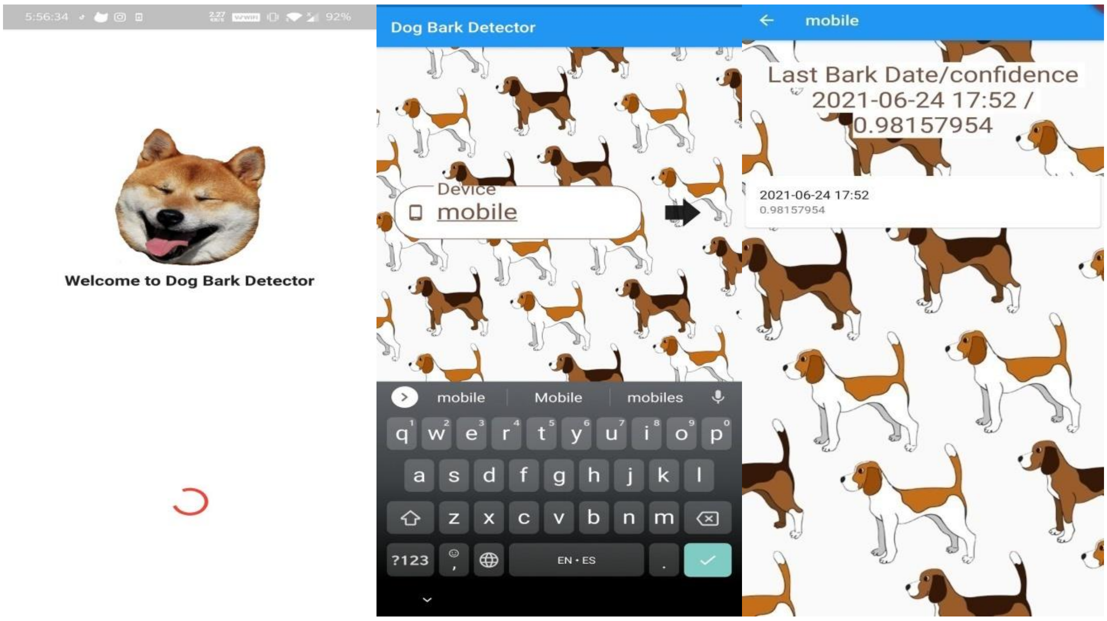
Figure 5. Real-time Recording, Prediction, and Integration of Firebase

### d. How the Code Works

The program was written in Python 3, ran on Raspberry Pi's built-in IDE Thonny, and employs machine learning library TensorFlow [10]. TensorFlow has a relatively higher training time compared to other libraries such as Pytorch [11] and Scikit-learn, but has lower prediction execution time and lower memory usage, therefore more adequate for Raspberry Pi. In our case, model training time is immaterial as all predictions are made with a pre-trained model, so TensorFlow stands out as the best choice.

## 3.4. Mobile App

### a. Each Screen (Screenshots)



Welcome to Dog Bark Detector

Dog Bark Detector

Device

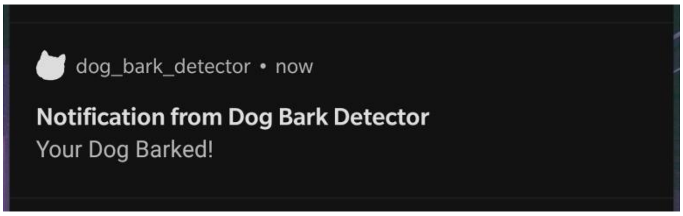
mobile

mobile

Last Bark Date/confidence
2021-06-24 17:52 /
0.98157954

2021-06-24 17:52
0.98157954

Figure 6. Screenshots of the screens



dog_bark_detector • now

Notification from Dog Bark Detector

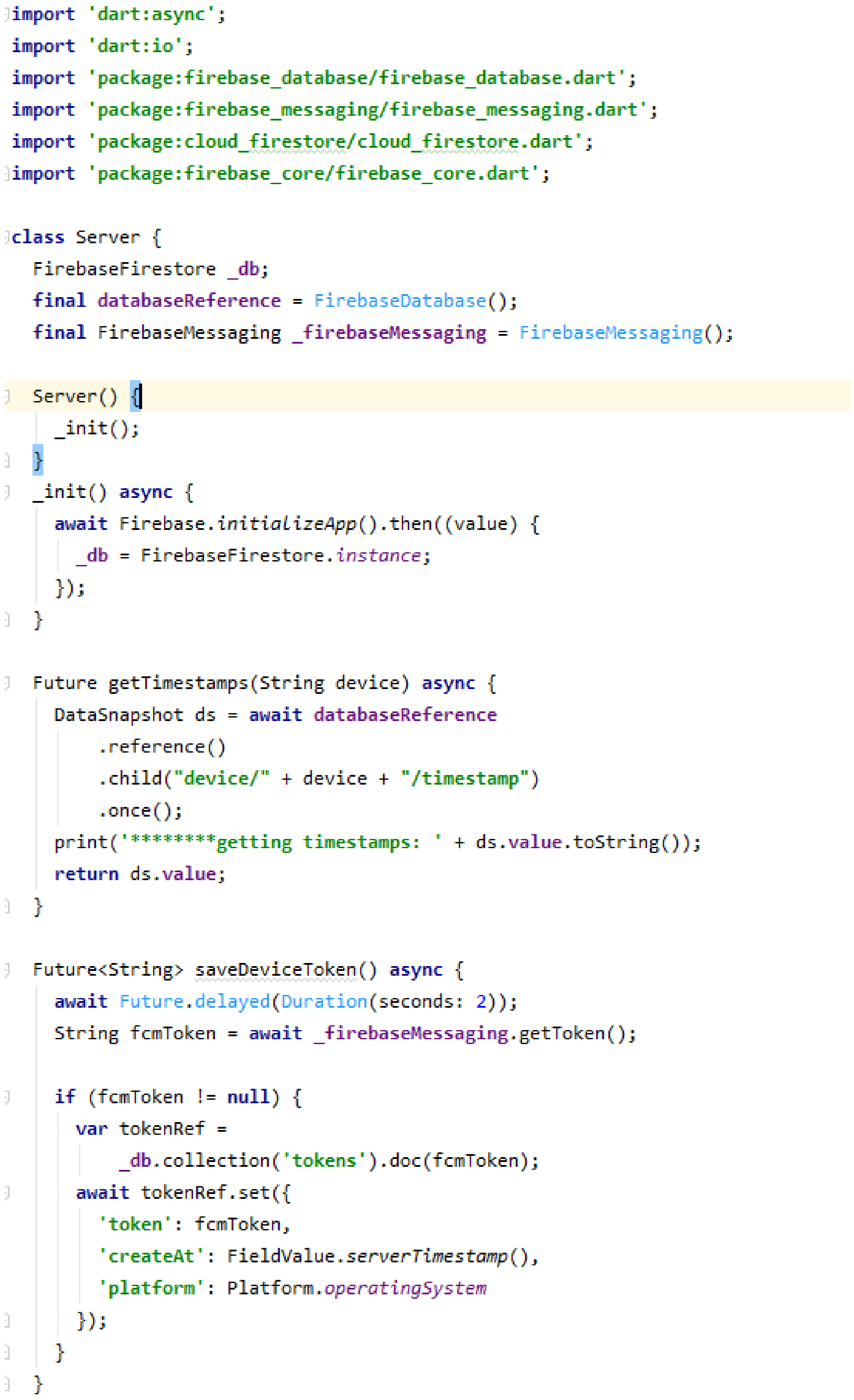
Your Dog Barked!

Figure 7. Notification Bar

### a. How Each Screen Works

As the mobile application boots up either through the notification or home screen icon, a 5-second long loading screen would be shown to indicate that the application is booting up. The user is then required to input the device's name to access the data. A page of recent history of dog barks is then shown along with the exact time each entry was recorded and the corresponding confidence score.

## b. Flutter Code



```
import 'dart:async';
import 'dart:io';
import 'package:firebase_database/firebase_database.dart';
import 'package:firebase_messaging/firebase_messaging.dart';
import 'package:cloud_firestore/cloud_firestore.dart';
import 'package:firebase_core/firebase_core.dart';

class Server {
  FirebaseFirestore _db;
  final databaseReference = FirebaseDatabase();
  final FirebaseMessaging _firebaseMessaging = FirebaseMessaging();

  Server() {
    _init();
  }
  _init() async {
    await Firebase.initializeApp().then((value) {
      _db = FirebaseFirestore.instance;
    });
  }

  Future getTimestamps(String device) async {
    DataSnapshot ds = await databaseReference
        .reference()
        .child("device/" + device + "/timestamp")
        .once();
    print('********getting timestamps: ' + ds.value.toString());
    return ds.value;
  }

  Future<String> saveDeviceToken() async {
    await Future.delayed(Duration(seconds: 2));
    String fcmToken = await _firebaseMessaging.getToken();

    if (fcmToken != null) {
      var tokenRef =
          _db.collection('tokens').doc(fcmToken);
      await tokenRef.set({
        'token': fcmToken,
        'createAt': FieldValue.serverTimestamp(),
        'platform': Platform.operatingSystem
      });
    }
  }
```

# 4. EXPERIMENT

## 4.1. Experiment 1

With the aim of improving the accuracy and precision of the model, we performed an experiment on prediction accuracy with varying dataset sizes. 10 trials are performed on three different models trained with 200, 400, and 800 sound files each for 10 categories of sounds. With each trial, we played a two-minute long sound file of solely dog barks, and recorded the percentage of times the machine predicted correctly.

Table 1. Experiments 4.1 Results

| | Mean Prediction Accuracy With 200 Files | Mean Confidence Score With 200 Files | Mean Prediction Accuracy With 400 Files | Mean Confidence Score With 400 Files | Mean Prediction Accuracy With 800 Files | Mean Confidence Score With 800 Files |
|---|---|---|---|---|---|---|
| Trial | | | | | | |
| 1 | 0.6898790124 | 0.8424729302 | 0.7559623831 | 0.8314935261 | 0.7631219106 | 0.7884614562 |
| 2 | 0.5912350912 | 0.7129731386 | 0.7066988174 | 0.8999849948 | 0.7921745436 | 0.8605691466 |
| 3 | 0.6581090928 | 0.8561823995 | 0.5864704471 | 0.8511170761 | 0.8597018772 | 0.7721071216 |
| 4 | 0.7362699784 | 0.7654463363 | 0.7819248494 | 0.7761451834 | 0.8241789856 | 0.8606500193 |
| 5 | 0.6383840589 | 0.9582189809 | 0.8404772978 | 0.7871010207 | 0.9029941285 | 0.8178276164 |
| 6 | 0.4983447521 | 0.8827262971 | 0.6831437616 | 0.8652125737 | 0.8908187336 | 0.8019187893 |
| 7 | 0.5379288024 | 0.6957023572 | 0.8202670667 | 0.8943896996 | 0.8373232305 | 0.9108755031 |
| 8 | 0.5818455893 | 0.8942314855 | 0.7184482631 | 0.7464330883 | 0.8736614863 | 0.8825037595 |
| 9 | 0.5860662396 | 0.7900477937 | 0.6779488965 | 0.9123124168 | 0.7590682019 | 0.7547074734 |
| 10 | 0.5147576467 | 0.8228336562 | 0.7335328985 | 0.8974635624 | 0.7108356222 | 0.9080990634 |

Though training the model with a larger dataset provides substantially higher accuracy, there was little to no change in the confidence score. By introducing more well-labeled data, we have significantly improved the accuracy of the model but with diminishing returns.

## 4.2. Experiment 2

After determining that a larger dataset provides better accuracy, we conducted another experiment to see if the number of categories of sounds affect the prediction accuracy and confidence. The setup of this experiment follows the first experiment, with the difference of varying categories instead of dataset sizes. We trained all three models with 800 sound files for each category but with varying 2, 5, 10 categories.

Table 2. Experiments 4.2 Results

| | Mean Prediction Accuracy With 2 Categories | Mean Confidence Score With 2 Categories | Mean Prediction Accuracy With 5 Categories | Mean Confidence Score With 5 Categories | Mean Prediction Accuracy With 10 Categories | Mean Confidence Score With 10 Files |
|---|---|---|---|---|---|---|
| Trial | | | | | | |
| 1 | 0.8855852903 | 0.8885125829 | 0.8390956366 | 0.7043971628 | 0.8556732618 | 0.7331123179 |
| 2 | 0.9042318007 | 0.9562397262 | 0.8286099065 | 0.7183265602 | 0.7935607238 | 0.7191559346 |
| 3 | 0.7224606445 | 0.8487778692 | 0.8523785396 | 0.8008522239 | 0.8251472019 | 0.7119563546 |
| 4 | 0.8844939738 | 0.9451752134 | 0.7721262744 | 0.8124913883 | 0.8443776904 | 0.6221453087 |
| 5 | 0.8281247652 | 0.8292085486 | 0.8791382616 | 0.7974484484 | 0.7656544446 | 0.6704566329 |
| 6 | 0.8409141153 | 0.8064941977 | 0.7566873349 | 0.7993017992 | 0.7845480411 | 0.8042016261 |
| 7 | 0.8972567475 | 0.8709239151 | 0.8998339055 | 0.7089790011 | 0.8675134809 | 0.7083710241 |
| 8 | 0.7063986176 | 0.7768236345 | 0.8594534284 | 0.748812237 | 0.8251472019 | 0.6794634922 |
| 9 | 0.8023372542 | 0.8553321055 | 0.7638194263 | 0.7413257356 | 0.7707758818 | 0.7280253879 |
| 10 | 0.7796276999 | 0.8458663361 | 0.7305136769 | 0.7044689465 | 0.8567961605 | 0.6494426147 |

The introduction of new categories affected the mean confidence of predictions due to more similarities between more categories, but the similarities were not indistinguishable, and were unable to affect the prediction accuracy.

### 4.3. Analysis

Through trials of experimentation, we were able to assess the overall effects of varying dataset size and categories. The increase of data in the dataset resulted in an overall improvement in prediction accuracy, which is reasonable as all of the data we used were accurately labeled. With the increase of categories, however, was able to negatively affect the prediction confidence, though not able to significantly affect prediction accuracy. The decrease in prediction confidence could be caused by the similarities between each type of sound, but the sounds are not similar enough to affect the actual prediction results.

## 5. RELATED WORK

Along the same lines of work, Mesaros, A. et al [13] utilize TUT Acoustic Scenes database to train models with the ability of sound event detection, specifically environmental acoustic scenes. In comparison, their data training employs an unsupervised Gaussian mixture model, while we utilized a supervised neural network as our dataset was manually labeled. In order to produce an accurate model through their method, it is safe to assume that a larger dataset is needed. In a scenario where training data is not abundant, our method of supervised training would be preferable, though with the downside of time-consuming manual labeling.

Adavanne, S. et al [14] investigate the difference between monaural and binaural sound event detection by studying different binaural features. The paper explores the substantial improvement in prediction accuracy along with the ability of classifying polyphonic sound events. Their method combines convolutional neural network [15] with recurrent neural network, producing a model that requires minimal supervision on pre-processing and secures high accuracy in polyphonic sound events. In our case, having the ability of polyphonic detection would be beneficial as the overlapping of sounds is a common occurrence in communities.

Wang, Y. [16] explores the advantages of transfer learning with a rather incomplete, poorly labeled dataset. Wang, Y. concluded that linear softmax pooling produces the best performing model through Convolutional Neural Network in the absence of a large, well-labeled dataset. Transfer learning, as discussed in the paper, would require the source task to have over 50 times the amount of data of the target task to achieve a successful application. The implementation of transfer learning, though proven to be useful, would not be realistic in our scenario as we have access to a rather large, well-labeled dataset.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we proposed an operational system that encompasses efficiency, accuracy to a sizable problem regarding the well being of community residents. The Dog Bark Detector allows users to be notified of the current ambience of their home environment as a form of automated system. The integration of a real-time database, real-time predictions and cloud messaging allows the user to receive notifications with minimal delay. Through trials of experiment, we were able to prove the high prediction accuracy of our program by finding the most efficient combination of training features. We found that, though with diminishing returns, simply increasing the size of the dataset with well-labeled data increases the accuracy of predictions. We also explored the impact of adding additional categories during model training, and came to a conclusion that adding categories of prediction results decreases the confidence score of prediction and its effects towards prediction accuracy are negligible.

The system has much to improve on, such as its prediction accuracy and practicality. The current model was trained through a dataset provided by UrbanSound8K. Despite having over 8732 sound files, only a small fraction of the sound files were dog barks as the dataset covered 10 categories ranging from “Jackhammer” to “Siren”. This significantly limits the amount of soundfiles going towards the training of specifically dog barks. One way to improve the accuracy of the predictions is to include more dog barks sound files in the dataset.

These limitations could be solved through either searching for a larger labeled dataset to produce a model that predicts with minimal errors, or include a verification segment where the program only sends notifications to its user when multiple predictions appear to be dog barks in a short span of time.

In K-12 educational settings, the program can be modified to strengthen the connection between teachers and administration and solve the in-class emergencies. For example, when a physical altercation takes place in class, since teachers are not allowed to touch the students, they need to let the class administration know about the situation. In this case, teachers can label the assistance that are needed to stop a physical altercation as “help one”. When the program predicts that the sound file signifies “help one”, the information of classroom number and teacher’s name will be sent to the school office. When the class administration receives the notification, the school secretary or principal can come over and stop the fight. In the same way, medical emergencies can be labeled as “help two”. When it happens, the teacher can perform CPR while reaching out to the administration to help call the ambulance.

## REFERENCES

[1] Jégh-Czinege, Nikolett, TamásFaragó, and PéterPongrácz. "A bark of its own kind–the acoustics of ‘annoying’ dog barks suggests a specific attention-evoking effect for humans." Bioacoustics 29.2 (2020): 210-225.

[2] “HOA Noise Rules: Complaining About Neighbor's Party Noise: HOAM.” HOA Management, 10 Dec. 2020, www.hoamanagement.com/hoa-noise-rules/.

[3] Singh, Narendra, and Subhash C. Davar. "Noise pollution-sources, effects and control." Journal of Human Ecology 16.3 (2004): 181-187.

[4] Yang, Honggang. "The disputing process: An ethnographic study of a homeowners association." Mediation Quarterly 13.2 (1995): 99-113.

[5] Hammer, Monica S., Tracy K. Swinburn, and Richard L. Neitzel. "Environmental noise pollution in the United States: developing an effective public health response." Environmental health perspectives 122.2 (2014): 115-119.

[6] Casey, Michael. "General sound classification and similarity in MPEG-7." Organised Sound 6.2 (2001): 153-164.

[7] Moroney, Laurence. "Firebase cloud messaging." The Definitive Guide to Firebase. Apress, Berkeley, CA, 2017. 163-188.

[8] Mandelbaum, Amit, and Daphna Weinshall. "Distance-based confidence score for neural network classifiers." arXiv preprint arXiv:1709.09844 (2017).

[9] Garg, Srishti, et al. "Urban Sound Classification Using Convolutional Neural Network Model." IOP Conference Series: Materials Science and Engineering. Vol. 1099. No. 1. IOP Publishing, 2021.

[10] Dillon, Joshua V., et al. "Tensorflow distributions." arXiv preprint arXiv:1711.10604 (2017).

[11] Paszke, Adam, et al. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32 (2019): 8026-8037.

[12] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[13] Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. "TUT database for acoustic scene classification and sound event detection." 2016 24th European Signal Processing Conference (EUSIPCO). IEEE, 2016.

[14] Adavanne, Sharath, and Tuomas Virtanen. "A report on sound event detection with different binaural features." arXiv preprint arXiv:1710.02997 (2017).

[15] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." 2017 International Conference on Engineering and Technology (ICET). Ieee, 2017.

[16] Wang, Yun. "Polyphonic sound event detection with weak labeling." PhD thesis (2018).

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN ML-BASED MEMORY LEAK DETECTION SCHEME FOR NETWORK DEVICES

Minghui Wang, Jiangxuan Xie, Xinan Yang, Xiangqiao Ao

AI Research Institute, H3C Technology Co., Ltd, China

## *ABSTRACT*

*The network is very important to the normal operation of all aspects of society and economy, and the memory leak of network device is a software failure that seriously damages the stability of the system. Some common memory checking tools are not suitable for network devices that are running online, so the operation staff can only constantly monitor the memory usage and infer from experience, which has been proved to be inefficient and unreliable. In this paper we proposed a novel memory leak detection method for network devices based on Machine learning. It first eliminates the impact of large-scale resource table entries on the memory utilization. Then, by analyzing its monotonicity and computing the correlation coefficient with the memory leak sequence sets pre constructed by simulation, the memory leak fault can be found in time. The simulation experiments show that the scheme is computationally efficient and the precision rate is close to 100%, it works well in the actual network environment, and has excellent performance.*

## *KEYWORDS*

*Memory leak, Resource table entry utilization, Correlation coefficient, Time Sequence monotonicity, Machine Learning.*

## 1. INTRODUCTION

The network is essential to the normal operation of all aspects of society and economy. Many large-scale network systems require 7x24h uninterrupted operation. Network quality is largely reflected in the stability and data forwarding ability of network device under long-term work. These network devices include switches, routers, firewalls, etc. They may have some software or hardware failures from time to time, and memory leakage is one of the serious software problems. It seriously damages the stability of the system, even leads to system crash or device restart.

Generally speaking, memory leak refers to that the application program does not release the memory in time after using it, and the memory can no longer be used by other applications. In severe cases, the memory will gradually run out, other applications will not be able to apply for the memory, and the system will crash eventually.

Memory leak detection includes static analysis and dynamic monitoring. Static analysis [1-4] is usually lexical, grammatical checking and type analysis for source code. The dynamic monitoring method [5-7] is to insert memory leak detection code at the location of memory operation to track memory usage, and report detailed information when a leak occurs. These methods require complex resource managements and modifications to the original application code. Although these software can effectively detect memory leaks, they have a relatively large run-time overhead and tend to reduce the efficiency of the system [8].Therefore, these types of check are generally disabled for the officially released software version, especially for network devices.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 27-40, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111403

For network devices, a common way to find memory leaks is to constantly monitor memory usage. If you find that the memory usage of some application is increasing and significantly exceeds the normal allowed level, consider the possibility of memory leaks. However, this method depends heavily on the experience of the operation staff, which means low efficiency and poor reliability.

In response to the urgent need for online device memory leak detection, this paper designs a memory leak detection method for network device based on Machine Learning(ML). By periodically monitoring the number of resource table entries (ARP, Route, MAC, ACL, etc.) and memory utilization of network devices, we can judge whether there is a potential memory leak and evaluate the time to reach the memory alarm threshold. Furthermore, the detailed memory usage information of the device can be obtained through the rule engine, which is convenient for the operation staff to complete the specific problem diagnosis. This method can detect the device memory leakage fault conveniently and quickly. It has the characteristics of accuracy, high efficiency, strong practicability and wide application range. This feature has been applied to Seer Analyzer—the network analyzer product of H3C company, and has achieved excellent results.

The rest of this paper is organized as follows. In section II some commonly used memory leak detection methods for network device are briefly described. Section III proposes a memory leak detection algorithm based on machine learning. Section IV provides some follow-up processing. Finally, experiment results and conclusions are presented in sections V and VI, respectively.

## 2. BACKGROUND AND PROBLEM

As mentioned before, for network devices, a common way to find memory leaks is to constantly monitor memory usage. When memory occupancy is abnormal (for example, the memory exceeds the alarm threshold, and the memory size increases abnormally), the operation staff check the occupancy of each memory block, analyze the allocation and release of suspicious memory blocks and their relationship with related applications, so as to speculate whether the application program has memory leakage fault.

However, this judgment process may cause false positives in memory leak alarming. For example, the increase in the usage of some memory blocks is normal. Only by combining the memory growth rate and the memory footprint with continuous observation can an accurate judgment be made. Therefore, the use of this method depends heavily on the experience of the operation staff, which means low efficiency and poor reliability.

In addition, some memory leaks are very slow and require long time (even months) to monitor and analysis, which greatly increases the difficulty of finding anomalies. Sometimes the continuous increasing of a small amount of memory may not be a problem. For example, the syslog data generated during system operation will be stored in the memory file, causing the memory to increase gradually until the file is written into the Flash or Disk periodically.

The testing department often uses the following methods to check for memory leaks: repeatedly delivering and deleting configurations, repeatedly delivering and deleting resource table entries, leaving the configuration or entries unchanged for a period of time to see if there is a big memory change. Among them, "repeatedly delivering and deleting configurations" rarely appear on actual network devices, so it is not considered here.

Due to the time limit of the software version plan, it is almost impossible to test the device for too long. However memory leaks are often related to some configurations or scenarios that require long-time observation to be found [9]. Considering this, memory leaks are not supposed to be found completely during the test phase. Consequently, it is necessary to monitor the device memory, use big data and ML technologies to store and analyze the data, and detect device memory leaks online.

## 3. Memory Leak Detection Algorithm Based on Machine Learning

In this section, we will introduce the memory leak detection algorithm based on machine learning. First of all, we exclude the influence of large-scale resource table entries on the memory utilization of devices and get a new time series called M' sequence. We preliminarily judge whether there is the risk of memory leakage in this new sequence according to its monotonic rising property. Moreover, we construct several memory leak sequence sets by simulation and compare the M' sequence with it. If the average correlation coefficient obtained is greater than a specified threshold value, it is further confirmed as an outlier.

The proposed scheme is as follows:

### 3.1. Eliminate the impact of large-scale resource entries on the device memory utilization, and get a new time series

As we know, most of the large-scale resource table entries on the device, such as ARP table entries, Routing table entries, MAC table entries and ACL table entries, are dynamically delivered and deleted, and the large size of these table has a great impact on the device memory utilization.

Therefore, without considering these resource table entries, it doesn't make sense to just monitor the changes in the device memory utilization. Only when the impact of these large-scale resource entries is excluded can the risk of possible memory leaks be exposed.

Network device, generally refers to switch or router, its large-scale resource table entries mainly include ARP entries, Routing entries, MAC entries, ACL entries, etc.

Let the utilization rate of ARP, Route, MAC and ACL entries be a%, b%, c% and d%, respectively. Here, the utilization rate of ARP entries represents the number of ARP entries currently used divided by the maximum ARP entry specification, and similar definitions for other entries. As follows show:

$$\begin{aligned} a\% &= arp\_num\_used \,/\, arp\_num\_total \\ b\% &= route\_num\_used \,/\, route\_num\_total \\ c\% &= mac\_num\_used \,/\, mac\_num\_total \\ d\% &= acl\_num\_used \,/\, acl\_num\_total \end{aligned} \quad (1)$$

Further, let $arp\_size$ represent the size of the memory occupied by each ARP entry, then $arp\_size\_total = arp\_size \times arp\_num\_total$ represents the memory size occupied by the ARP maximum entry specification, and similar definitions for $route\_size\_total$, $mac\_size\_total$, and $arp\_size\_total$.

Now, we can calculate the weighted sum of the utilization of these resource entries:

$$x\%=w_1\times a\%+w_2\times b\%+w_3\times c\%+w_4\times d\% \quad (2)$$

Where $w_1$, $w_2$, $w_3$, and $w_4$ are defined as follows represents the weight values of each resource table entry’s impact on device memory:

$$\begin{aligned}
w_1 &= arp\_size\_total/(arp\_size\_total+route\_size\_total+mac\_size\_total+acl\_size\_total) \\
w_2 &= route\_size\_total/(arp\_size\_total+route\_size\_total+mac\_size\_total+acl\_size\_total) \\
w_3 &= mac\_size\_total/(arp\_size\_total+route\_size\_total+mac\_size\_total+acl\_size\_total) \\
w_4 &= acl\_size\_total/(arp\_size\_total+route\_size\_total+mac\_size\_total+acl\_size\_total)
\end{aligned} \quad (3)$$

Using gRPC[10](Google Remote Procedure Call) technology, network devices periodically send the utilization rate of memory and the resource tables to the analyzer, each of these data forms a time series.

$$\begin{aligned}
M &= ( m_1, \cdots, m_n) \\
A &= ( a_1, \cdots, a_n) \\
B &= ( b_1, \cdots, b_n) \\
C &= ( c_1, \cdots, c_n) \\
D &= ( d_1, \cdots, d_n)
\end{aligned} \quad (4)$$

Where M, A, B, C, D represent the time series composed of device memory utilization, ARP entry utilization, Route entry utilization, MAC entry utilization and ACL entry utilization.

From *A*, *B*, *C*, *D*, we can calculate their weighted sum at each time point to get a new time sequence.

$$R=(r_1, \cdots, r_n), \; r_i=w_1 \times a_i+w_2 \times b_i+w_3 \times c_i+w_4 \times d_i, \; i=1, \cdots, n \quad (5)$$

Where the value range of $r_i$ is [0,100]. We divide the range into 500 equal-length cells with fixed length of 0.2. Then the number of elements of the sequence *R* falling between each cell is calculated and the cell with the largest number is regarded as cell *u*.

Suppose there are *p* elements in *R* whose values belong to cell *u* and the corresponding time points are $t_1'$, $t_2'$,...$t_p'$. Correspondingly, the sequence *M* takes values at these time points to form a new sequence：

$$M'= (m_1', \cdots, m_p') \quad (6)$$

Next, we will examine the monotonicity of the new sequence and its correlation with some known memory leak sequences to figure out whether the device has memory leakage.

It should be noted that ARP, Route, MAC and ACL table we selected here are the most common large-scale resource tables. If there are other more large-scale resource tables that may seriously affect the memory utilization, they also need to be considered here to participate in the calculation.

For network device, the weight values $w_1$, $w_2$, $w_3$, and $w_4$ barely change with the software version. Only when there are many significant changes in the processing of related resource table entries in the software version, their values need to be determined again. Experiments show that even if the values of these weights are slightly changed, the analysis results will not be affected.

### 3.2. Determine the monotonicity of the M' sequence

For the memory sequence *M'* obtained in the previous step, we know that the sum of the memory occupied by the large-scale resource table entries at each time point is approximately equal. In this section, we will judge the monotonicity of *M'* sequence by calculating the correlation coefficient between the *M'* sequence and its index sequence (*I*=(*1,2,3,...,p*)).

Regarding memory leakage, the memory utilization will keep rising with the passage of time without the impact of large-scale resource entries. Therefore, if the *M'* sequence follows the monotonous upward trend, it indicates that there may be a memory leak. The correlation coefficient between *M'* and its index sequence is calculated. Here we use the Spearman algorithm [11], which is a rank correlation algorithm and does not require the assumption of bivariate normal distribution, so it has a better effect than the Pearson algorithm in this case. Here, we set the threshold value of correlation coefficient as 0.9, above which represents the monotonic rising trend of *M'* series. See the following experimental section for details.

In order to eliminate random interference, only *M'* sequence with the memory increment (which is the difference value between the first item and the last item) exceeding the specified threshold (e.g. 30Mbytes) will be regarded as the potential memory leakage sequence. Consequently, the interference of a small increase of memory caused by syslog and diagnostic information of each software module will be successfully eliminated. In other words, for the sequence ( $m_1'$, …, $m_p'$), it is required to satisfy $(m_p'-m_1')\%*total\_mem > 30M$ , where *total_mem* is the total memory of the device.

To sum up, if the sequence meets two conditions: time monotonic growth and the memory increment exceeds the specified threshold, then the possibility of memory leak of the device can be preliminarily judged.

### 3.3. Calculate the average correlation coefficient between M' and simulation-constructed memory leak sequence sets

In this section, we first assume some memory leak scenarios and construct a set of corresponding time series. Their *M'* sequences are calculated respectively to form a simulated memory leak sequence set.

Due to the monotonous rising trend of memory leakage *M'* sequence, it has strong correlation with simulated memory leak sequence; while the normal *M'* sequence shows random fluctuation trend, and the correlation with simulated memory leak sequence is very weak. We calculate the correlation coefficients of the *M'* sequence and a pre-generated simulated memory leak sequence by Spearman algorithm and take the average. If the average correlation coefficient is greater than 0.9, it indicates that the *M'* sequence has a memory leak. Please see the following experimental section for details.

The processing flow chart of the scheme described in this section is shown below：

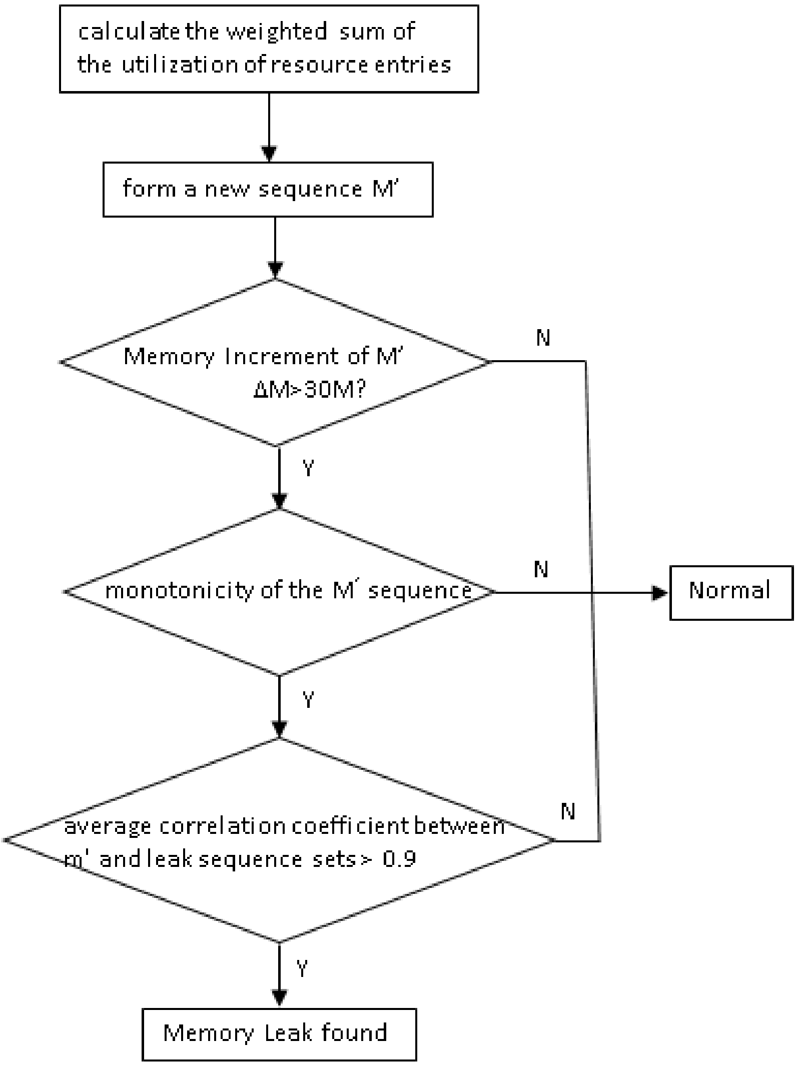

calculate the weighted sum of the utilization of resource entries
form a new sequence M'
Memory Increment of M' ΔM>30M?
N
Y
monotonicity of the M' sequence
N
Normal
Y
average correlation coefficient between m' and leak sequence sets > 0.9
N
Y
Memory Leak found

Figure 1. Flow chart of the scheme

## 4. FURTHER PROCESSING

In this section, we first assume some memory leak scenarios and construct a set of corresponding time series. Their *M'* sequences are calculated respectively

If it is determined that a memory leak has occurred in a certain *M'* sequence, by calculating the memory increment *M_Diff* of the sequence and the corresponding time difference *T_Diff*, we can get the average growth rate of memory leakage $v=M\_diff/T\_diff$ . Consequently, we can estimate the time when the memory reaches the alarm threshold which help the operation staff to reasonably arrange the time for fault handling. Specifically, the corresponding estimated time is: $Est\_T=total\_mem\times(alarm\%-curr\%)/v$ where *alarm*% represents the memory alarm threshold, *curr*% represents the current memory utilization, and *total_mem* represents the total memory size of the device, then the memory growth space left to reach the alarm threshold is $total\_mem\times(alarm\%-curr\%)$ .

After detecting the memory abnormality, further memory diagnostic tools can be used. As known, the rule engine [12-13] is a component embedded in the application which can separate business rules from the business code and use pre-defined semantic specifications to implement these separated rules. Given input data, the rule engine perform evaluate rules and make decisions. With the aid of rule engine, we use NETCONF [10] (Network Configuration Protocol) to interactively obtain the detailed information of each memory block from the device, it can assist the operation staff to further confirm the problem and find the faulty application module.

## 5. Simulation Experiments and Results

In the simulation experiments, we collected the data of the utilization rate of each resource table entry and the device memory utilization rate periodically. We collected the data every 5 minutes over 3 months and the total data point is $288 \times 30 \times 3 = 25920$. Here 3 scenarios are involved: the normal situation without memory leaks, the memory leaks of ARP resource entries, and the random memory leaks of an unknown software module. We use python to complete the simulations.

The following are the graphs of simulated device memory in three scenarios：

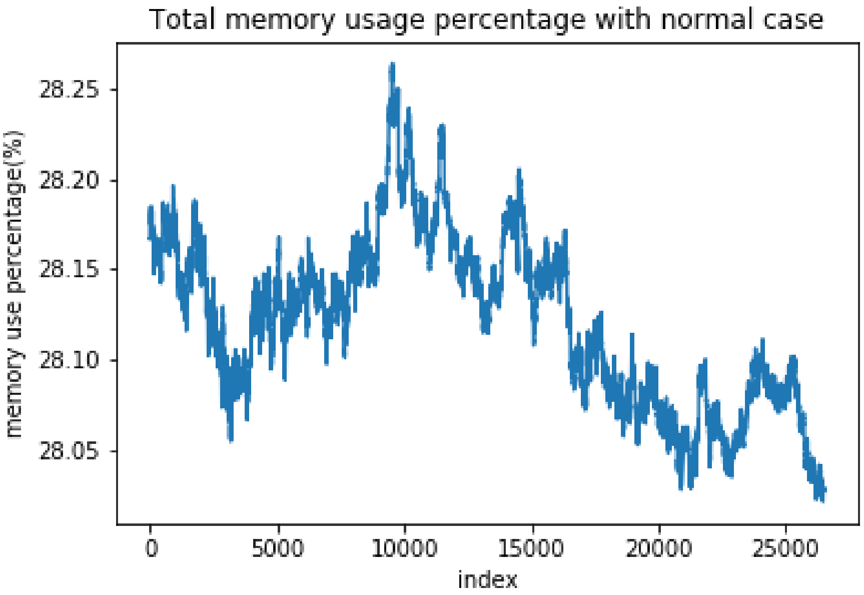


Figure 2. Total memory usage percentage with normal Case

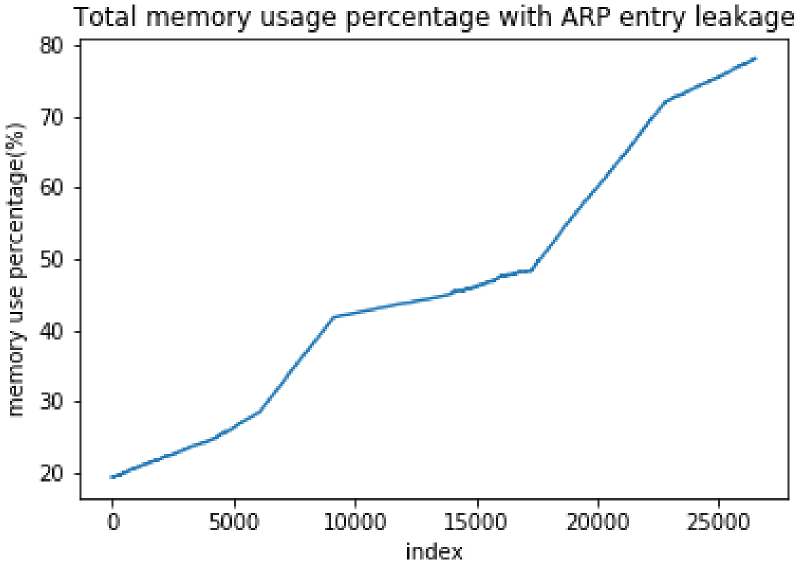


Figure 3. Total memory usage percentage with ARP Leakage

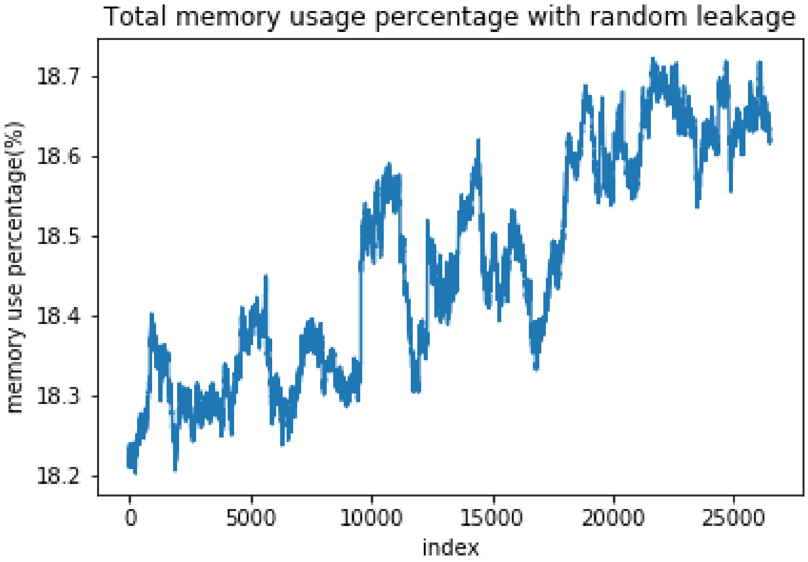


Figure 4. Total memory usage percentage with Random Leakage

The graphics of their corresponding *M*' sequences are shown below：

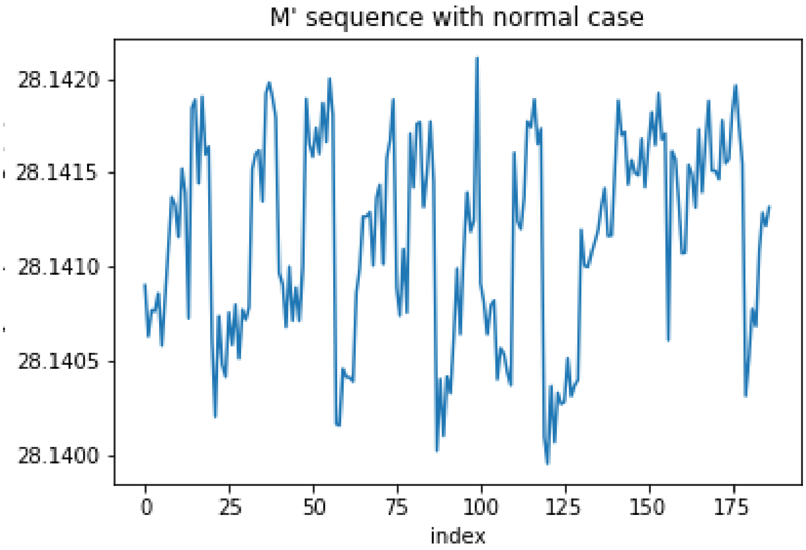


Figure 5. M’ sequence with Normal Case

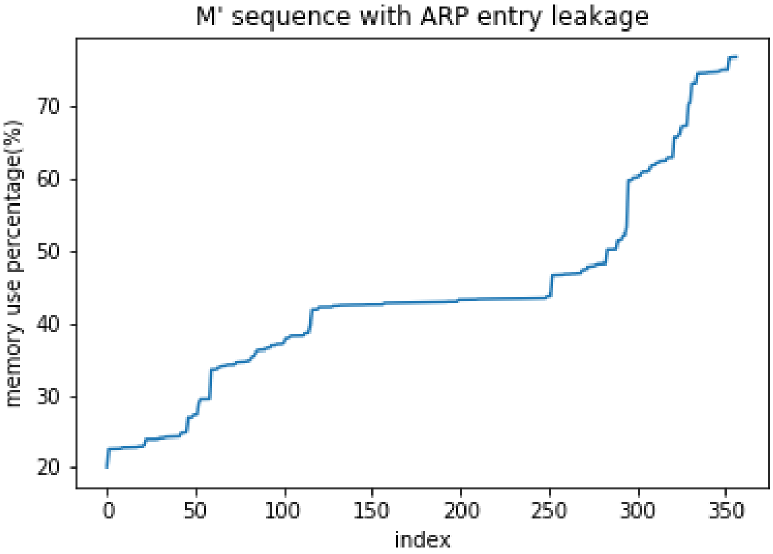


Figure 6. M’ sequence with ARP entry Leakage

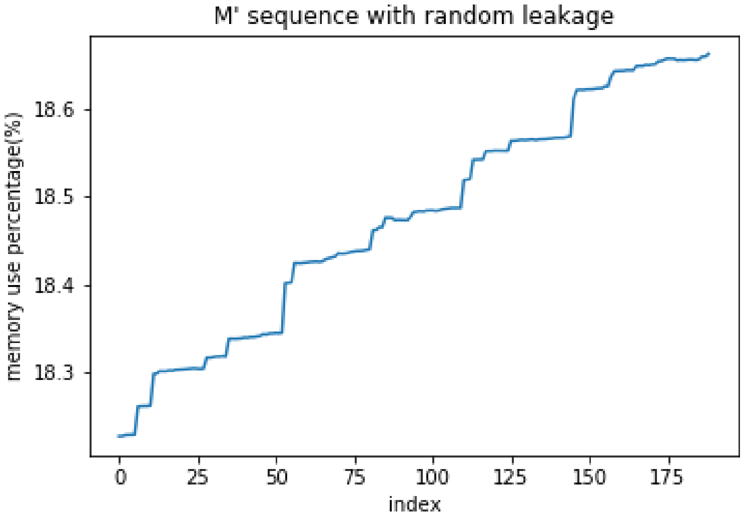


Figure 7. M’ sequence with Random Leakage

Above figures clearly show that the memory utilization of normal *M'* sequence fluctuates slightly and the curve is random; the memory utilization rate of resource leak *M'* sequence changes greatly, showing an obvious monotonous upward trend; while the memory utilization rate of random leak *M'* sequence changes little and the overall trend is monotonous increasing.

Next, we simulated 500 sequences for each of the three scenarios and calculated their *M'* sequences. Related statistics including the memory increment of the sequence, the time correlation, and the average correlation coefficient (using Spearman algorithm) with the simulated memory leak sequences. They are recorded in table 1-3, but only the first 20 pieces of data are shown.

Table 1. Test Sequences with Normal case

| | **Memory Difference** | **Corr. with index sequ.** | **Corr. with Resource leak sequ.** | **Corr. with random leak sequ.** |
|---|---|---|---|---|
| 1 | 291248 | 0.158175215 | 0.155201666 | 0.145934015 |
| 2 | 301214 | -0.019766941 | -0.010182252 | 0.018031106 |
| 3 | 260485 | 0.109665269 | 0.092678231 | 0.08981839 |
| 4 | 291945 | 0.112001576 | 0.165478754 | 0.174005038 |
| 5 | 290986 | -0.094858308 | -0.113018566 | -0.11931812 |
| 6 | 291624 | -0.252379809 | -0.22554809 | -0.225589206 |
| 7 | 284432 | -0.086634815 | -0.049478148 | -0.048171725 |
| 8 | 344911 | 0.161756547 | 0.154608783 | 0.147779058 |
| 9 | 251442 | -0.302939227 | -0.208654013 | -0.210560823 |
| 10 | 295035 | -0.060953324 | -0.041223962 | -0.049084516 |
| 11 | 276335 | 0.02740196 | 0.096420279 | 0.097325497 |
| 12 | 318737 | 0.37203624 | 0.319781804 | 0.313736033 |
| 13 | 280037 | 0.073193108 | 0.084716381 | 0.089715752 |
| 14 | 295190 | -0.013274524 | -0.040550019 | -0.044731653 |

| | | | | |
|---|---|---|---|---|
| 15 | 295725 | 0.007098027 | 0.041243151 | 0.046395803 |
| 16 | 278126 | -0.011541039 | 0.030574502 | 0.024268845 |
| 17 | 267995 | -0.09952762 | -0.112263921 | -0.115108548 |
| 18 | 298624 | 0.189605496 | 0.035233616 | 0.024033341 |
| 19 | 278243 | 0.143553709 | 0.108807238 | 0.119960642 |
| 20 | 273391 | 0.090146257 | 0.128837797 | 0.129857563 |

Table 2. Test Sequences with Resource Leak case

| | **Memory Difference** | **Corr. with index sequ.** | **Corr. with Resource leak sequ.** | **Corr. with random leak sequ.** |
|---|---|---|---|---|
| 1 | 2789276999 | 0.999998 | 0.999956 | 0.978997 |
| 2 | 2013235537 | 0.999989 | 0.999934 | 0.979096 |
| 3 | 2620491023 | 0.999892 | 0.999828 | 0.978786 |
| 4 | 2807450398 | 0.999999 | 0.999963 | 0.978941 |
| 5 | 2809413438 | 0.999998 | 0.999962 | 0.978945 |
| 6 | 1521223836 | 0.999991 | 0.999952 | 0.978938 |
| 7 | 2202391094 | 0.999991 | 0.999952 | 0.978986 |
| 8 | 972756363 | 0.999964 | 0.999926 | 0.979024 |
| 9 | 2441715706 | 0.999932 | 0.999958 | 0.979001 |
| 10 | 2838246478 | 0.999997 | 0.999961 | 0.978919 |
| 11 | 2436498700 | 0.999996 | 0.999963 | 0.979145 |
| 12 | 2007848751 | 0.999994 | 0.999957 | 0.978956 |
| 13 | 2658185467 | 0.999999 | 0.999962 | 0.978915 |
| 14 | 735720744 | 0.99997 | 0.999922 | 0.978947 |
| 15 | 1690241815 | 0.999993 | 0.999914 | 0.978523 |
| 16 | 2668271684 | 0.99999 | 0.999932 | 0.978854 |
| 17 | 968070171 | 0.999991 | 0.999956 | 0.9791 |
| 18 | 3425955833 | 0.999997 | 0.999959 | 0.978903 |
| 19 | 3530875222 | 0.999998 | 0.99997 | 0.979072 |
| 20 | 1531155291 | 0.999995 | 0.999957 | 0.978918 |

Table 3. Test Sequences with Random Leak case

| | **Memory Difference** | **Corr. with index sequ.** | **Corr. with Resource leak sequ.** | **Corr. with random leak sequ.** |
|---|---|---|---|---|
| 1 | 39257102 | 0.999806 | 0.979739374 | 0.97879 |
| 2 | 38617581 | 0.999781 | 0.979660463 | 0.978735 |
| 3 | 36811484 | 0.999386 | 0.979299968 | 0.978057 |
| 4 | 40963370 | 0.999746 | 0.97967272 | 0.978652 |
| 5 | 39622467 | 0.997962 | 0.977857124 | 0.976249 |
| 6 | 40131717 | 0.999658 | 0.979572898 | 0.978668 |
| 7 | 37434253 | 0.999524 | 0.979181516 | 0.978452 |
| 8 | 35191411 | 0.999252 | 0.979071952 | 0.978153 |
| 9 | 37450356 | 0.999353 | 0.979354171 | 0.978422 |

| | | | | |
|---|---|---|---|---|
| 10 | 39239786 | 0.998732 | 0.978647139 | 0.977538 |
| 11 | 38587988 | 0.998366 | 0.978069213 | 0.977219 |
| 12 | 39072316 | 0.99942 | 0.978734961 | 0.977913 |
| 13 | 24657491 | 0.99928 | 0.979058344 | 0.978111 |
| 14 | 37803860 | 0.99986 | 0.979789375 | 0.978773 |
| 15 | 37887778 | 0.999711 | 0.979627497 | 0.978657 |
| 16 | 38422163 | 0.999802 | 0.979595805 | 0.978723 |
| 17 | 36231575 | 0.999263 | 0.978949737 | 0.977975 |
| 18 | 36120339 | 0.99965 | 0.97959303 | 0.978576 |
| 19 | 39842950 | 0.999748 | 0.979516662 | 0.978655 |
| 20 | 38039417 | 0.997934 | 0.977837991 | 0.97655 |

Follow the process shown in Figure 1, the memory increment of normal $M'$ sequence is less than 30M, so they are all judged as normal data. For the ARP table entry memory leak $M'$ sequence, its time correlation reaches 99.9% on average(greater than 0.9), the average correlation coefficient with the resource memory leak sequence set is greater than 99% and the average correlation coefficient with the random memory leak sequence set is greater than 97 %. Apart from these, the 497 out of 500 sequences show large memory increment which are beyond 30M and are judged as leaked data. For the random memory leak $M'$ sequence, its time correlation reaches 99.9% on average(greater than 0.9), the average correlation coefficient with the resource memory leak sequence set is greater than 97% and the average correlation coefficient with the random memory leak sequence set is greater than 97%. Slightly lower proportion than the ARP table entry memory leak $M'$ sequence, 452 out of 500 sequences are judged as leaked data. No misjudgment has been made in all scenarios which shows out 100% precision in automatic check for memory leak detection.

It should be pointed out that some memory leaks are very slow and do not exceed 30M in 3 months. At this increasing trend, only 360M of memory will be leaked in 3 years, which has little impact on the system, and it is not necessary to deal with it.

It can be seen from the above simulation experiments that the algorithm described in this article has outstanding advantages of high efficiency and high accuracy. This solution has been applied in the Seer Analyzer product of H3C Company which can detect various memory leaks and achieve excellent performance.

## 6. CONCLUSION

Memory leaks in network devices make some memory unavailable for subsequent use which may cause service failure or even system crash in severe cases. Commonly used manual detection methods are inefficient and error-prone, and cannot perform online and real-time detection for a large number of network devices. The memory detection scheme proposed in this paper could automatically check the device memory occupancy rate and combines the resource occupancy information of various table entries of the device to judge for memory leaks. Then estimate time to reach memory alarm threshold will be predicted and the rule engine will be used to get more detailed diagnostic information. The simulation experiments show that the scheme is computationally efficient and the precision rate is close to 100%, which solves this problem well and is of great practical significance.

This solution has been applied in the Seer Analyzer product of H3C Company, we also intend to apply it to the network devices of other companies. It should be noted that different manufacturers have different table item sizes, so the weight values of the tables should be chosen carefully, they may have an important impact on the accuracy of the algorithm. The table item sizes even change with the software version, so it is very useful to automatically find the changes and adaptively adjust parameters.

## REFERENCES

[1] John Regehr & Nathan Cooprider & Will Archer, (2006) “Efficient type and memory safety for tiny embedded systems,” In: Proc of the 3rd workshop on Programming languages and operating systems: linguistic support for modern operating systems, San Jose, California,pp.22–22.

[2] Thomas A. Henzinger & Ranjit Jhala & Rupak Majumdar & Marco A.A.Sanvido,(2004) “ Extreme Model Checking,” In Verification: Theory and Practice, Lecture Notes in Computer Science 2772, Springer-Verlag, pp.332–358.

[3] Hu Yan & Gong Yu-chang & Sun Wei-feng & Zhao Zhen-xi (2008) “Hybrid Static Method for Memory Leak Detection,” Journal of Chinese Computer Systems, Vol29,pp.1935–1939.

[4] Gong Yu-chang & Hu Yan & Zhang Ye & Zhao Zhen-xi,(2009) “A static memoryleak detection method for binary programs, ” Journal of University of Science and Technology of China,Vol39,pp. 189–195.

[5] K. Chen & J.-B. Chen.(2007) Aspect-Based Instrumentation for Locating Memory Leaks in Java Programs. In IEEE International Conference on Computers, Software & Applications (COMPSAC), pages 23–28.

[6] M. Jump & K. S. McKinley.(2007) Cork: Dynamic memory leak detection for garbage-collected languages. In ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL), pages 31–38.

[7] C. Jung & S. Lee & E. Raman & S. Pande.(2014) Automated Memory Leak Detection for Production Use. In International Conference on Software Engineering (ICSE), pages 825–836.

[8] G. Xu & A. Rountev.(2013) Precise Memory Leak Detection for Java Software Using Container Profiling. ACM Transactions on Software Engineering and Methodology, 22(3):17:1–17:28.

[9] M. D. Bond & K. S. McKinley.(2009) Leak Pruning. In ACM Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pages 277–288.

[10] Zohaib Latif & Kashif Sharif & Fan Li & Yu Wang.(2020) A comprehensive survey of interface protocols for software defined networks, Journal of Network and Computer Applications.

[11] R. L. Iman & W. J. Conover.(1982) A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics: Simulation and Computation, 11(3):311–334.

[12] Yong H. Lee & Suk I.Yo0. (1995) A Rete-based Integration of Forward and Backward Chaining Inferences, ISIC. Page(s): 611 – 616.

[13] Sun, Y. & Wu, T.Y. & Zhao, G. & Guizani, M., (2015) Efficient rule engine for smart building systems. IEEE Trans. Comput. 64, 1658–1669. https://doi.org/10.1109/TC.2014. 2345385.

## AUTHORS

**Minghui Wang** received his B.Sc degree and PhD degree in Mathematics from The Peking University, in 1996 and 2001, respectively. Currently, is a Senior Engineer with Institute for Artificial Intelligence, H3C Co, Ltd. His research interests include IP network communication, data mining and AI.

**Jiangxuan Xie** received his B.Sc degree in Statistics from The University of Hong Kong, Hong Kong, China in 2018. Currently, he is a Machine Learning Engineer for H3C, AI Research Institute. His current research interests include areas of time series analysis, data mining, big-data algorithms.

**Yang Xin'an** received his B.Sc. degree in electronics and information engineering from Nanjing University of Science and Technology, Nanjing, China, in 1999 . Currently, he is a Senior Engineer with Institute for Artificial Intelligence, H3C Co, Ltd. His current research interests include AI, BigData, Network communication.

**Xiangqiao Ao** received his B.Sc. degree in Computer Science from Zhejiang University, Hangzhou, China, in 2001. He has been working in Huawei and H3C since 2001, focusing on IP network communication. Currently he is in charge of the AI Institute of H3C. His research interests include IP network communication, AI-based IP network and Autonomous-Driving-Network.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN INTELLIGENT DRONE SYSTEM TO AUTOMATE THE AVOIDANCE OF COLLISON USING AI AND COMPUTER VISION TECHNIQUES

Steven Zhang[1] and Yu Sun[2]

[1]Crean Lutheran High School, Irvine, CA 92618
[2]California State Polytechnic University, Pomona, CA, 91768

## ABSTRACT

*People love to fly drones, but unfortunately many end up crashing or losing them. As the technology of flying drones improves, more people are getting involved. With the number of users increasing, people find that flying drones with sensors is safer because it can automatically avoid problems, but such drones are expensive. This paper describes an inexpensive UAV (unmanned aerial vehicle) system that eliminates the need for sensors and uses only the camera to avoid collisions. This program helps avoid drone crashes and losses. We used the Tello Education drone as our testing drone, which is only outfitted with a camera. Using the camera feed and transmitting that data to the program, the program will then give commands to the drone to avoid collisions.*

## KEYWORDS

*Machine Learning, Electrical Engineering, Computer Vision, Drone*

## 1. INTRODUCTION

With the continuous development of science and technology, UAV (unmanned aerial vehicle) technology is also getting recognized by new users. [1, 2, 3, 4] With the increasing number of people participating in UAV, there are many companies specializing in manufacturing this technology. [5] But there's a problem. UAVs are small, flexible, and crucially, pilotless, which means they're vulnerable to damage or accidents during flight. [6] Using sensors, drones can automatically avoid collisions, but this feature comes with problems. First, it is expensive to build, and second, it is heavier. We designed a UAV system to effectively avoid these problems. Our UAV system uses a camera to scan the surrounding space, then the background processing system is used to calculate the most suitable solutions for directional movement for collision avoidance. This makes our UAV system affordable to build and reduces unnecessary weight.

Using existing technology, combined with our group design program, the camera scans the environment. Data is sent to the terminal to calculate and determine obstacles and the best directional path and automatic correction to avoid the obstacles. Our system doesn't need a new camera or censors, only the UAV's own camera and a program to avoid collisions and compute directional movement. Many of today's drones have automatic avoidance technology, but this usually requires special sensors. This means their costs are higher and weights are heavier. For any UAV, weight is an important consideration for flight range, and lighter models are generally more economical as well.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 41-48, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111404

We searched for a drone platform that would allow us to code and have good flying control and visual feedback. After some research, we ended up choosing the Tello model as out project drone. Our choice and method of using a positive camera feed to catch and process images was inspired by the DJI Phantom 4 Pro drone. [8] Our drone can also make decisions to avoid objects. Our ETCollision drone has many useful features. First, there is no need to intervene to have the drone avoid obstacles. Second, there are no extra accessories to be installed on the drone itself. Therefore, the drone flight time won't be affected, since there is no extra weight. Third, in the future we hope we can add faster processing as a feature. For example, perhaps we can improve the drone's reaction time with a photo process. This will allow for a quicker response when the drone needs to avoid objects. Moreover, due to having a quicker response time, we can also improve the program to make it to do more complicated stunts and further reduce the possibility of crashes. Overall, we believe the ETCollision is a program that will reduce the possibility of crashing and has potential for improvements over time as well.

In two application scenarios, we demonstrated how the above combination of features increases the UAV's performance. First, we conducted a comprehensive case study on the evolution of the Tello drone, which allowed us to have a precise understanding of this model's movement and performance, especially when navigating the drone around objects quickly and smoothly without crashing. All of our data was calculated by the centimeter, and was double-checked by GPS to make sure that we had a drone up and running that would get the job done. [9, 10] What's more, we coded the drone with automatic flight control, which means if there was wind or other conditions that affected the drone, the drone will automatically adjust back into its original flight path. Therefore, the drone can operate under harsh conditions. For example, if a wind is blowing from north to south, the drone will automatically exert energy to make sure the wind does not carry it away from the desired flight path or into obstacles. Second, we analyzed the evolution of data from each time we allowed updates from the drone. We always consulted our data to make sure there were no bugs or errors at any given time. All in all, our ETCollision Tello drone has excellent performance and precise GPS data to ensure that it can securely get the job done at all times.

## 2. CHALLENGES

In order to develop an inexpensive UAV (unmanned aerial vehicle) system that eliminates the need for sensors and uses only the camera to avoid collisions, a few challenges were identified as follows.

### 2.1. Challenge 1: Choosing a Platform and Drone to Use

Our first challenge was to decide on a platform to design our Collision project. Out of several platforms, we decided on Python, since it is easy to use compared to other program platforms. Therefore, after deciding on Python, will needed to find a drone that could let us use it to code it while still achieving excellent flight control. We chose Tello, since it offers advantages that other drones do not. For example, Tello has a GPS system and an excellent flight time of 15 minutes so we don't have to worry as much about batteries. What's more, Tello also has an HD camera that allows 30 fps feedback from the drone with very little delay.

### 2.2. Challenge 2: Setting Up Positive Video Feedback from the Drone

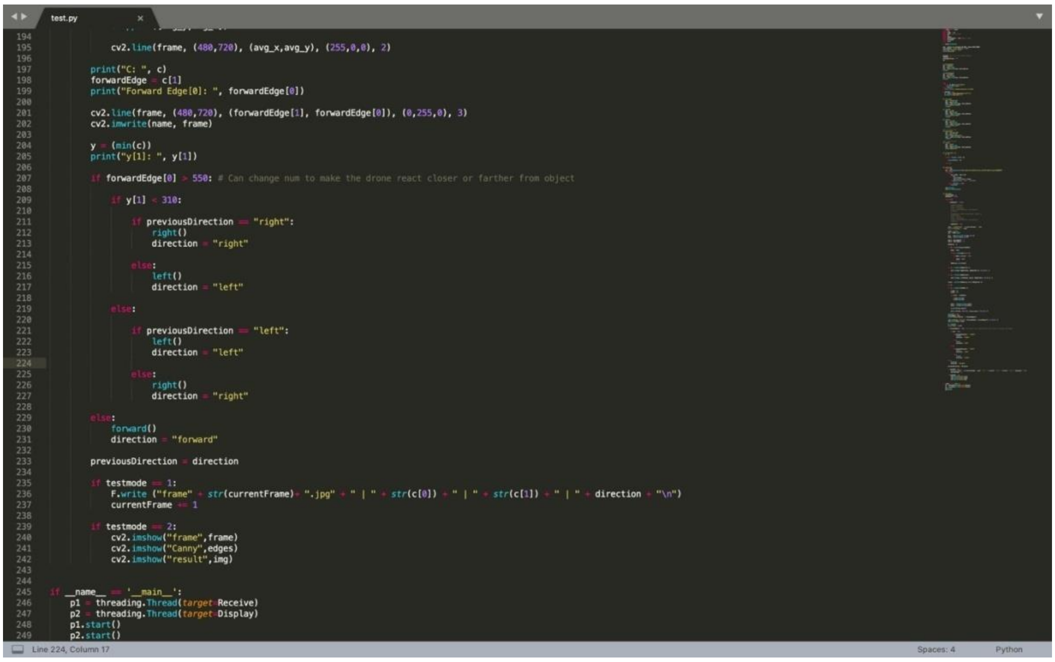
Our second challenge was to have a positive drone feed. Initially, our code was allowing the drone to detect objects, but the feedback form the drone had a 10-second delay. Therefore, the drone had a slow response time and an inefficient flight time. At this time, all the drone was doing was hovering and waiting for the video feedback to get processed. We solved this by using a positive video feedback form the drone, and allowing a window pop up on our computer to speed up the drone feedback and allow us to see the drone's flight path.

### 2.3. Challenge 2: Getting The Drone on the Market

The last challenge was buying a domain name, making a website, and entering a competition. It was difficult to choose a domain name that no one else had registered. We wanted to come up with a name that's easy to remember and find in web searches. We also had difficulty recording the screen, since we could not find software that allowed us to record our screen and our voice at the same time. However, we solved this problem by using different software. For the domain, we used "etcollision.com" so it would be easy to remember.

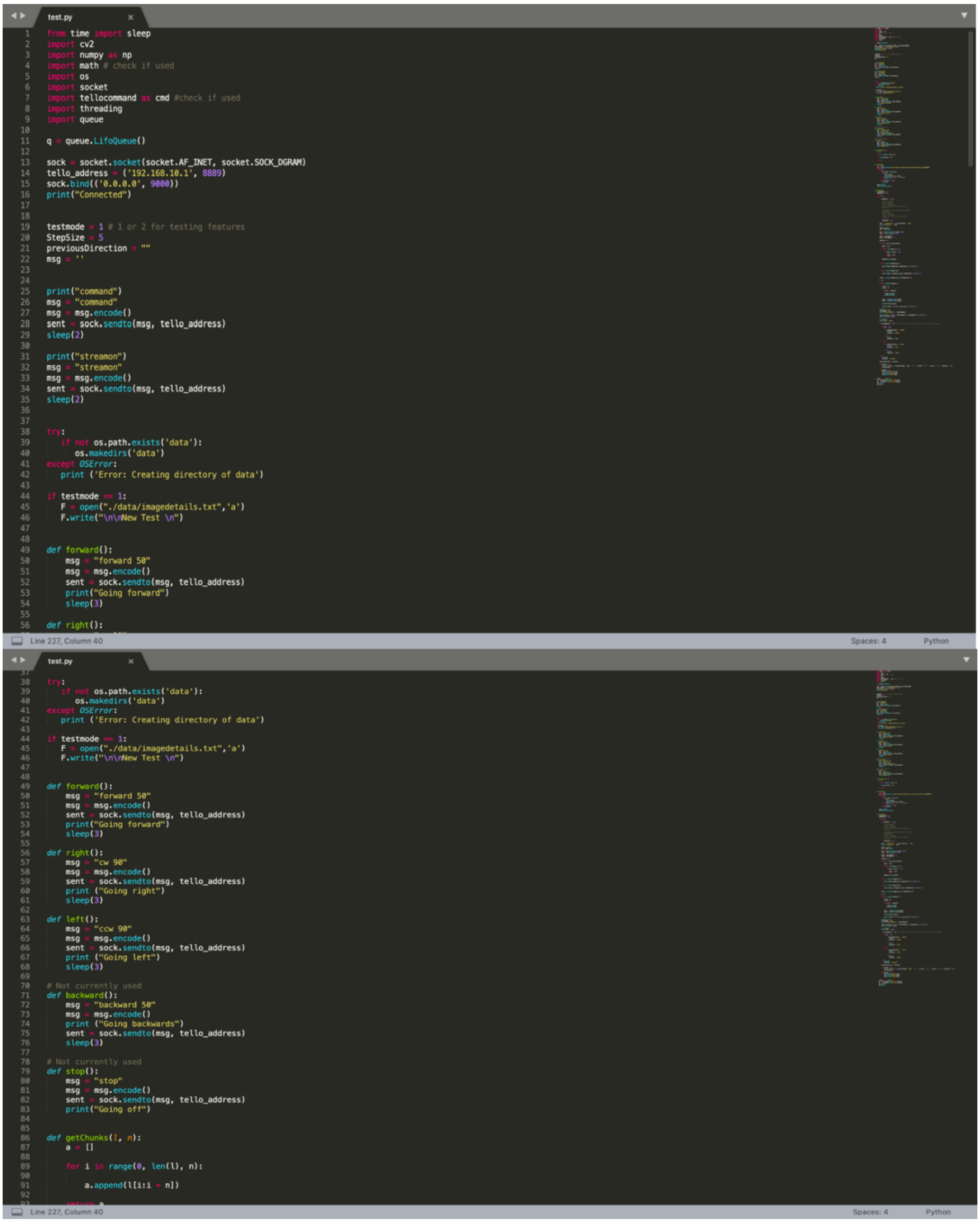
## 3. SOLUTION

Our ETCollision drone is a system that allows the drone to process video intake to avoid objects on its own. Using Python, we wrote code to process feedback form the drone. The code was scaled in centimeters, which allowed precise feedback from the drone. This special coding allowed us to see the drone's flight movement. The drone provided video feedback like a normal drone does, and it ran through a special code that allowed it to detect and process objects within its airspace. Therefore, the drone calculated varying flight courses to avoid objects. Moreover, while the drone was in the air, the command prop launched and opened a window that allowed us to see the drone's video feedback with very little delay. We could also give commands or press the emergency stop button when needed. All the drone's movements were tested to be done in a split second to make sure it would work without delay. In the future, we hope we can set up a larger data processor so the drone could remember where it has been, have faster reaction and maneuver times and navigate more smoothly and quickly.



```python
            cv2.line(frame, (480,720), (avg_x,avg_y), (255,0,0), 2)

        print("C: ", c)
        forwardEdge = c[1]
        print("Forward Edge[0]: ", forwardEdge[0])

        cv2.line(frame, (480,720), (forwardEdge[1], forwardEdge[0]), (0,255,0), 3)
        cv2.imwrite(name, frame)

        y = (min(c))
        print("y[1]: ", y[1])

        if forwardEdge[0] > 550: # Can change num to make the drone react closer or farther from object

            if y[1] < 310:

                if previousDirection == "right":
                    right()
                    direction = "right"

                else:
                    left()
                    direction = "left"

            else:

                if previousDirection == "left":
                    left()
                    direction = "left"

                else:
                    right()
                    direction = "right"

        else:
            forward()
            direction = "forward"

        previousDirection = direction

        if testmode == 1:
            F.write ("frame" + str(currentFrame)+ ".jpg" + " | " + str(c[0]) + " | " + str(c[1]) + " | " + direction + "\n")
            currentFrame += 1

        if testmode == 2:
            cv2.imshow("frame",frame)
            cv2.imshow("Canny",edges)
            cv2.imshow("result",img)


if __name__ == '__main__':
    p1 = threading.Thread(target=Receive)
    p2 = threading.Thread(target=Display)
    p1.start()
    p2.start()
```

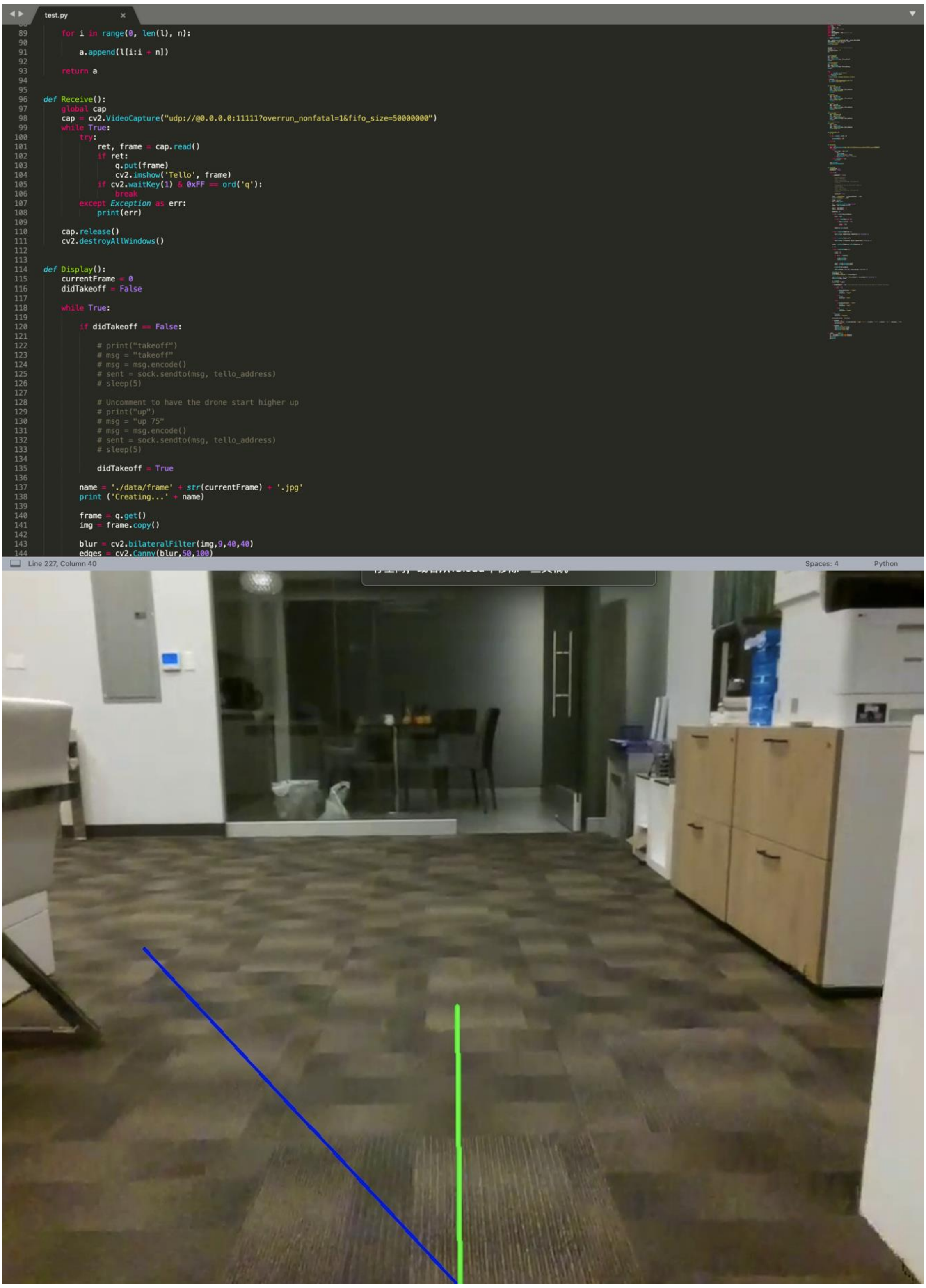
Figure 1a. Code segments



```python
from time import sleep
import cv2
import numpy as np
import math # check if used
import os
import socket
import tellocommand as cmd #check if used
import threading
import queue

q = queue.LifoQueue()

sock = socket.socket(socket.AF_INET, socket.SOCK_DGRAM)
tello_address = ('192.168.10.1', 8889)
sock.bind(('0.0.0.0', 9000))
print("Connected")


testmode = 1 # 1 or 2 for testing features
StepSize = 5
previousDirection = ""
msg = ''


print("command")
msg = "command"
msg = msg.encode()
sent = sock.sendto(msg, tello_address)
sleep(2)

print("streamon")
msg = "streamon"
msg = msg.encode()
sent = sock.sendto(msg, tello_address)
sleep(2)


try:
    if not os.path.exists('data'):
        os.makedirs('data')
except OSError:
    print ('Error: Creating directory of data')

if testmode == 1:
    F = open("./data/imagedetails.txt",'a')
    F.write("\n\nNew Test \n")


def forward():
     msg = "forward 50"
     msg = msg.encode()
     sent = sock.sendto(msg, tello_address)
     print("Going forward")
     sleep(3)

def right():
```

```python
try:
    if not os.path.exists('data'):
        os.makedirs('data')
except OSError:
    print ('Error: Creating directory of data')

if testmode == 1:
    F = open("./data/imagedetails.txt",'a')
    F.write("\n\nNew Test \n")


def forward():
     msg = "forward 50"
     msg = msg.encode()
     sent = sock.sendto(msg, tello_address)
     print("Going forward")
     sleep(3)

def right():
     msg = "cw 90"
     msg = msg.encode()
     sent = sock.sendto(msg, tello_address)
     print ("Going right")
     sleep(3)

def left():
     msg = "ccw 90"
     msg = msg.encode()
     sent = sock.sendto(msg, tello_address)
     print ("Going left")
     sleep(3)

# Not currently used
def backward():
     msg = "backward 50"
     msg = msg.encode()
     print ("Going backwards")
     sent = sock.sendto(msg, tello_address)
     sleep(3)

# Not currently used
def stop():
     msg = "stop"
     msg = msg.encode()
     sent = sock.sendto(msg, tello_address)
     print("Going off")


def getChunks(l, n):
     a = []

     for i in range(0, len(l), n):

         a.append(l[i:i + n])

     return a
```

Figure 1b. Code segments, continued



```
    for i in range(0, len(l), n):

        a.append(l[i:i + n])

    return a


def Receive():
    global cap
    cap = cv2.VideoCapture("udp://@0.0.0.0:11111?overrun_nonfatal=1&fifo_size=50000000")
    while True:
        try:
            ret, frame = cap.read()
            if ret:
                q.put(frame)
                cv2.imshow('Tello', frame)
            if cv2.waitKey(1) & 0xFF == ord('q'):
                break
        except Exception as err:
            print(err)

    cap.release()
    cv2.destroyAllWindows()


def Display():
    currentFrame = 0
    didTakeoff = False

    while True:

        if didTakeoff == False:

            # print("takeoff")
            # msg = "takeoff"
            # msg = msg.encode()
            # sent = sock.sendto(msg, tello_address)
            # sleep(5)

            # Uncomment to have the drone start higher up
            # print("up")
            # msg = "up 75"
            # msg = msg.encode()
            # sent = sock.sendto(msg, tello_address)
            # sleep(5)

            didTakeoff = True

        name = './data/frame' + str(currentFrame) + '.jpg'
        print ('Creating...' + name)

        frame = q.get()
        img = frame.copy()

        blur = cv2.bilateralFilter(img,9,40,40)
        edges = cv2.Canny(blur,50,100)
```

Figure 1c. Code segments and display image

1-9 import library (the video feedback from the drone will be imported to the library for later review)

25-45 load data (the data will load in a format that the system can process)

49-68 drone movement (After the program processes the movement it will start to control the drone to do the movement)

79-83 drone action (After the drone decides to make a move, all movements are ruled by the action code)

86 object detection (this is the process when the drone is processing the picture that is coming back from the video feedback)

114-242 display

245-end run all code

We decided to use Python for our platform code, since it is was the best coding platform for our drone. We did not add any additional system or hardware to do this, only add code to allow the program to process video feed from the drone for navigational purposes. See Figures 1a.-1c.

## 4. EXPERIMENT

At first, we did not have a positive video feed going to our computer to see what the drone was seeing. Therefore, we could not be sure if the drone was doing its job. We had to use a different command to allow the drone to process the video feed on its own to avoid objects. Moreover, we allowed the drone to make more precise movements without overreacting to objects. All in all, after lots of testing and coding, we were able to make the project work the way we designed it to.

We ran a percent test with our drone by directing it toward an object multiple times to calculate its passing percent and near misses. From this, we could get a percentage accuracy of its object detecting system. Moreover, we also changed the settings to make the drone more precise and flexible in certain situations, for example, varying weather conditions.

After we had conducted some experiments, we finally got an answer of 90 percent. Since our drone was relying on image processing, some of the objects could not be detected from the images alone. For example, some of the poles within our test area were not able to be picked up and processed by the drone’s video feedback. However, it was able to pick up the pole as a large shape. Moreover, it was unable to process clear glass or windows since the video feedback system can see through it. This caused the process system to think there was not an object present when at times there was glass present. Therefore, due to these situations, we only passed 90 percent of our experiment.

At first, we did not have a good video feed from our drone, which we had to fix. After doing that, we were able to move on to our second test, which was getting a percentage on avoiding objects. The problem we had with the drone was that it was not picking up some poles or clear glass. Therefore, we changed our settings on the drone to make it more precise. However, that also lowered the battery life down from the original 15 minutes of flight time. After the drone changed to high process mode, the flight time shrank to 10 minutes. In the future we hope to come up with new code to allow us to do the video feed process without using too much battery power.

## 5. Related Work

One related work is the self-recovery system used in DJI drones, which has a strong connection with the drone. The possibility of losing a connection is very small. However, they still developed a self-recovery mode to the drone to make the drone fly back to the original point of takeoff. Moreover, with this system, one can also set an altitude limit for its recovery flight. The drone will climb up to that height first, then fly back, which is a good system to make sure the drone doesn't hit anything on the way back. [11, 12]

Another related work is that DJI's drones all have average flight times of 30 minutes. Even in sport mode while at peak performance, their drones can still stay in the air for about at least 20 minutes. This is something we need to learn and study, since our drone had 15 minutes of flight time generally, and only 10 minutes of flight time while at perk performance. [13, 14]

There is also a low battery warning on DJI drones. However, since our drone doesn't have a controller, this information cannot be displayed. What we came up with to remedy this was to let the drone return by itself once its battery power hit a certain minimum. That way, we did not lose the drone or cause a crash. We also adjusted this feature further so we could also program the flight distance to make sure our drone could make it back every time. [15]

## 6. Conclusion and Future Work

We produced a special set of code to allow our drone to avoid objects in its path. To do this, we used video feedback from a camera that was already on the drone. We used Python as our primary coding platform, because it was the most compatible with our drone. Therefore, we could ensure the best performance of our code and drone at all times.

Our drone has few limitations. All a pilot might need would be to make sure to have a good battery and know how to start the drone. They must also know how to run the code and have quality video feedback showing on their computer screen. They also need to know how to land the drone and make an emergency landing in case the code has an error or the drone's video process program fails. Therefore, the limitations of our drone are few, since all these skills could be learned in under ten minutes.

One feature we hope to add is a self-recovery mode. This allows the drone to return to the place of takeoff in case of emergency. For example, if the drone loses connection with the computer, it can fly back by itself while the video feedback process system is still engaged to make sure it doesn't crash on the way back. There are lots of advantages for such a program. It could allow drones to fly out of sight and still make it back to the landing zone and can be used even if there are different signals jamming the drone's own signal. With such a self-recovery mode, our drone would not fall from a high attitude and be more likely to make it more home safely.

### References

[1] Lidynia, Chantal, Ralf Philipsen, and Martina Ziefle. "Droning on about drones—acceptance of and perceived barriers to drones in civil usage contexts." Advances in human factors in robots and unmanned systems. Springer, Cham, 2017. 317-329.

[2] Hendry, David. "“Drones Okay” Playground: Fun with Personal Drones." Designing Tech Policy.

[3] LaFay, Mark. Drones for dummies. John Wiley & Sons, 2015.

[4] Juniper, Adam. The Complete Guide to Drones: Whatever Your Budget. Wellfleet Press, 2016.

[5] Liu, Zhongli, et al. "Rise of mini-drones: Applications and issues." Proceedings of the 2015 Workshop on Privacy-Aware Mobile Computing. 2015.

[6] Vacek, Joseph J. "The next frontier in drone law: liability for cybersecurity negligence and data breaches for UAS operators." Campbell L. Rev. 39 (2017): 135.

[7] Wu, Wenhao. "React Native vs Flutter, Cross-platforms mobile application frameworks." (2018).

[8] Peppa, M. V., et al. "Photogrammetric assessment and comparison of DJI Phantom 4 pro and phantom 4 RTK small unmanned aircraft systems." ISPRS Geospatial Week 2019(2019).

[9] Gowda, Mahanth. "Bringing differential GPS to drones." Proceedings of the 3rd Workshop on Hot Topics in Wireless. 2016.

[10] Bo-tao, W. U., et al. "Testing and Analysis on differential GPS aerial drones technology." Journal of Yangtze River Scientific Research Institute 34.1 (2017): 142.

[11] Putch, A. N. D. Y. "Linear measurement accuracy of DJI drone platforms and photogrammetry." San Francisco: DroneDeploy(2017).

[12] Iqbal, Farkhund, et al. "Drone forensics: a case study on DJI phantom 4." 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2019.

[13] Salamh, Fahad E., Mohammad Meraj Mirza, and Umit Karabiyik. "UAV Forensic Analysis and Software Tools Assessment: DJI Phantom 4 and Matrice 210 as Case Studies." Electronics 10.6 (2021): 733.

[14] Xu, Fangqi, and Hideki Muneyoshi. "A Case Study of DJI, the Top Drone Maker in the World." Kindai Manag. Rev 5 (2017): 97-104.

[15] Yousef, Maryam, Farkhund Iqbal, and Mohammed Hussain. "Drone Forensics: A Detailed Analysis of Emerging DJI Models." 2020 11th International Conference on Information and Communication Systems (ICICS). IEEE, 2020.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# BABY CRY CLASSIFICATIONS USING DEEP LEARNING

Shane Grayson[1] and Wilson Zhu[2]

[1]Windward School, Los Angeles, USA
[2]Diamond Bar High School, Los Angeles, USA

## *ABSTRACT*

*New parents are frequently awakened by the cries of their newborn babies. Attempts to stop these cries sometimes result in increasingly louder cries. By first transforming these cries into waveforms, and then into sound spectrograms, the efficiencies and accuracies of different computer learning modules were tested: a support vector machine, a 2-layer neural network, and a long short-term memory model. Finally, an automatic sorter that categorizes each cry was developed. Using this method, it is possible to eliminate error and time wastage when trying to calm a baby. The results of testing the programs demonstrate a high accuracy rate for determining the source of a baby's cries. This program will enable parents to calm their crying babies in a shorter amount of time, giving them more peace of mind, and perhaps allowing them to get more sleep.*

## *KEYWORDS*

*Infant Cry, Deep Learning, Convolutional Neural Network, Audio Classification.*

## 1. INTRODUCTION

For the module, a combination of convolutional neural networks alongside other programs interpreting audio files was used. Computers did not start with this capability though, so in providing background for the topic, a brief history of the program that can understand written language, natural language processing (NLP), will be given. Natural Language Processing emanates from Noam Chomsky. In his book "Syntactic Structures" published in 1957, Dr. Chomsky theorized that if formatted in his style of grammar, called Phase Structure Grammar, computers would be able to understand human language [1]. In 1958, not long after "Syntactic Structures" was published, one of the first programming language processes, called LISP was released, and capable of giving textual responses in the form of a psychiatric technique called "reflection" [1].

Although high costs stopped most research regarding Natural Language Processing and AI in 1966, by 1980, NLP research was brought back with new, fresh ideas. Researchers abandoned the old ways of mixing statistics with linguistics because that method had not produced a computer program that even came close to carrying out conversations, instead, researchers adopted purely statistical models, even though the rise of early Machine Learning challenged these models; eventually leading to multiple high-level statistical NLP programs being created [1]. Breaking into the modern era, Deep Neural Network Learning has been the leading method for speech recognition, taking advantage of many different architectures, including support vector machines, maximum entropy models, neural networks, and Gaussian mixture models [2].

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 49-59, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111405

The growing number of architectures are now categorized into three different main groupings: Generative, Discriminative, and Hybrid deep learning architectures. Discriminative architectures can use visible data to determine patterns and characterize different aspects of the data [2]. Convolutional Neural Networks are just one of many examples of architecture in discriminative classification. These modules consist of a convolutional layer and a pooling layer, usually stacked, in order to form deep models [2]. They can enable computer programs to not only understand the language but to also understand the sentiment humans put behind that language. These models have been used in the healthcare field as NLP's which can deduce underlying meanings in forms filled out by patients to a high degree of accuracy, with similar methods being developed to be used on social media [3].

## 1.1. Existing Methods

These models can be applied in a broad array of situations, so it was difficult to narrow down all of the possible problems to the one that was settled on. Interest in developing an automated lab cry categorizer sprouted from the want to reduce the anxiety and sleep deprivation caused when new parents are forced to wake up at all hours of the night and care for their babies. Parents frequently have to test a multitude of ideas before one finally works. The effects of constantly trying to comfort a child are numerous and range from interrupting a parent's circadian rhythm (which, in particular, intensifies a mother's fatigue after the birthing process), to anxiety at the thought of harming one's child, and even depression [4]. While the added interaction may increase familiarity and trust in the parent, these side effects counteract that and can damage the parent-child relationship [4].

For these reasons, allocating the stress and work put on parents is a high priority. While there have been no groups using the same methods combined with this idea, there have been groups with a similar idea and separate execution, with their own pros and cons. For example, most recently a group participating in the 2019 10th International Conference on Computing, Communication and Networking Technologies tackled this issue by using Statistical Feature Extraction and Gaussian Mixture Models [5]. This team achieved an accuracy of 81.27%.

## 1.2. Method Used

As mentioned above, the most recent idea used Statistical Feature Extraction and Gaussian Mixture Models and was completed with an accuracy of 81.27% identifying a total of five different reasons for crying [5]. Gaussian Mixture Models are probability density functions that represent the weighted sums of component densities of a Gaussian [6]. While the use of Gaussian Mixture Models is a fascinating solution for determining differences between audio clips, and for inferring meaning through the different components of these cries, this method falls short on accuracy compared to other methods. That said, it does make up for this decreased accuracy by being extremely resistant to overfitting, allowing immense amounts of training to be done. Moreover, higher accuracy can be achieved using a support vector machine, while continuing to avoid the problem of overfitting. Overall, however, Gaussian Mixture Models are not the optimal method to use.

The methods used in this paper improves upon the accuracy used in previous approaches centering around Gaussian Mixture Models while avoiding issues like overfitting. This was achieved by first labeling the data into the reason for their cries: "belly_pain", "discomfort", "tired", "hungry", and "burping". After labeling all of the data, two built-in tensor flow tools were used to first decode the audio files and then transform them into waveforms, while simultaneously labeling them. This makes each audio file visible as a graph where some patterns are able to be seen; for instance, a pattern observable in the "hungry" waveforms is continuous and consistent pulse durations, which is seen as well in the “discomfort” labeled waveforms, just with much larger average crest and trough amplitudes. These both differ from the “tired” group, however, whose amplitude starts small comparatively and decreases alongside the wavelength as the cry continues. While the trends displayed by one group are difficult to notice, they are, in fact, discernible by the human eye. In order to enlarge these changes and highlight the trends previously seen, it was necessary to convert these waveforms into sound spectrograms on a logarithmic scale with labeled axes time and frequency.

Now, with the trends obvious, the three different models were trained, and the results were compared to determine the best way of interpreting this data. A 2-layer neural network, support vector machine, and long short-term memory model with results similar inaccuracy, but a loss rate much better from the support vector machine were used, leading to the conclusion that this method is best when categorizing baby cries by meaning.

### 1.3. Evaluations

Through multiple different applications, including a 2-layer neural network, support vector machine, and long short-term memory model, it can be concluded that the support vector machine is the most accurate method in determining the cause of a baby crying. Finishing with an accuracy of 86% compared to the other accuracies of 81%, it can be concluded that of the methods tested, the support vector machine is superior in classifying the sentiment behind a baby’s cries.

### 1.4. Paper Structure

The rest of this paper is organized in the following manner: Section 2 details the process and solution used to address the problem offered. Section 3 gives an overview of the results recorded during the evaluation of each program on the validation set as well as an analysis of the meaning behind these results. Finally, Section 4 offers a conclusion as well as a potential improvement upon the execution and ideas in future work.

## 2. OVERVIEW OF OUR APPROACH

In order to construct the processes described below, both Keras and TensorFlow were used as they comprise the largest python programming platform. Despite not being the most user-friendly software, TensorFlow allows for many different levels of customization, enabling the use of different methods. The model is split into three major processes: representing each data file as a waveform, transforming this waveform into a sound spectrogram, and finally using a series of different modules to train and test on these spectrograms.

The data files of babies crying were sourced from GitHub, where volunteer participants would download the Donate-a-cry application for either iOS or Android and submit an audio file of a baby crying along with the reason for why they were crying. These cleaned and filtered files are the ones used from GitHub as data samples for this paper [7]. Each file, when imported, was given a name matching the reason identified by the volunteers as the reason for crying: “belly_pain”, “discomfort”, “tired”, “hungry”, “burping”.

There was a total of 457 samples used. Of these 457 samples 360 were used for training, 40 for validation, and the other 57 were reserved for testing; this final accuracy is what was used to determine the overall effectiveness of the program. To transform these files into waveforms, they were first decoded using a TensorFlow built-in program. Once decoded the values here graphed with the y-axis representing time and x-axis representing amplitude. Originally the values for amplitude were given from -32768 to 32767, but were shrunk to fit between -1.2 and 1.2; once completed they looked as shown in Figure 1:

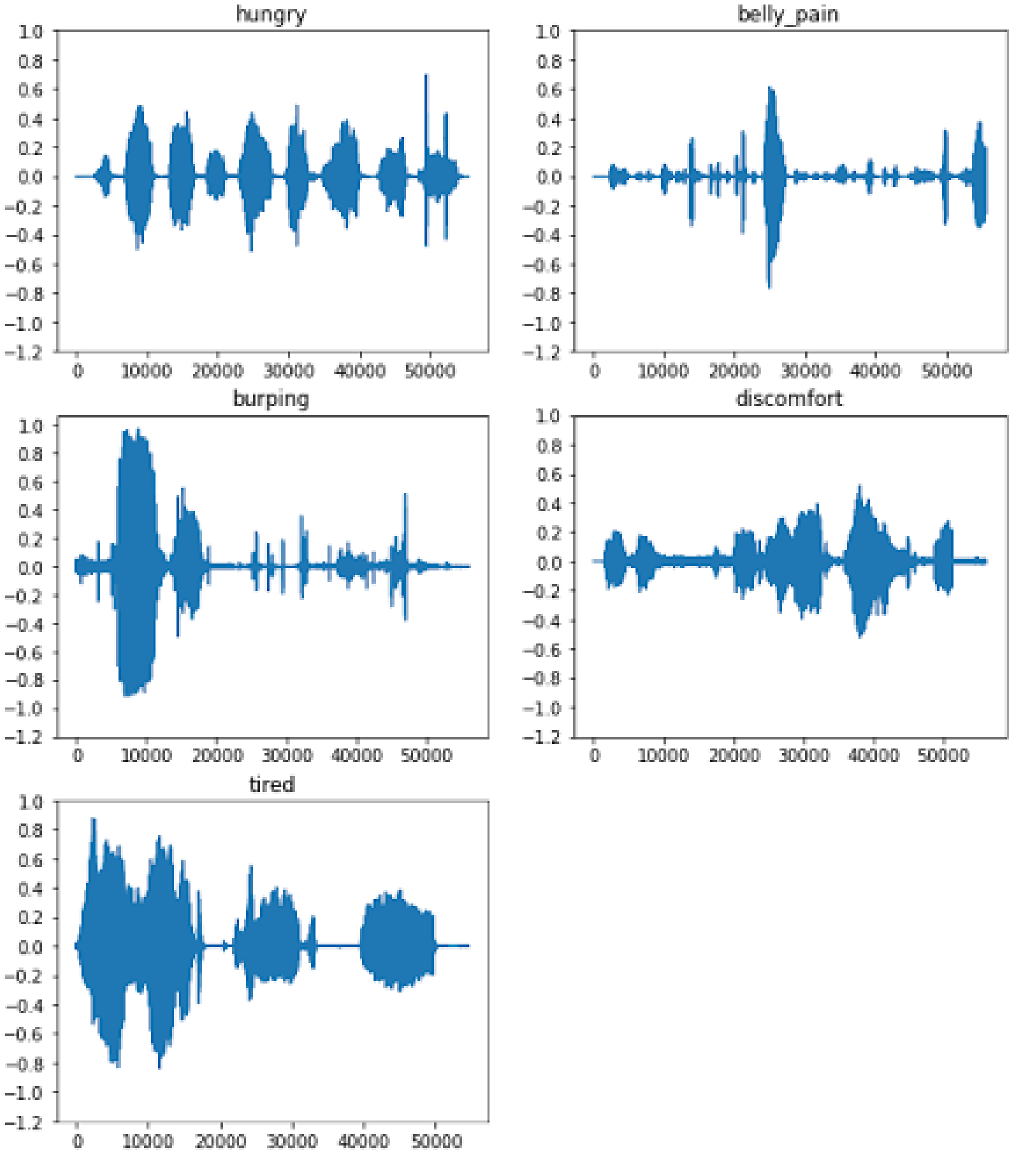


Figure 1. Examples of Data Waveforms

Afterward, each waveform of fewer than 60,000 samples was padded to be this length. They were then converted into spectrograms where the y-axis was time ranging from 0-60,000, the x-axis was measured in frequency from 0-120, and color represents amplitude with lighter color showing higher amplitudes and darker color showing lower amplitudes. To help accomplish this, the Fourier Transform mathematical concept has been used, which transforms the data from an audio signal into a frequency domain [8]. Figure 2 shows an example of these results.

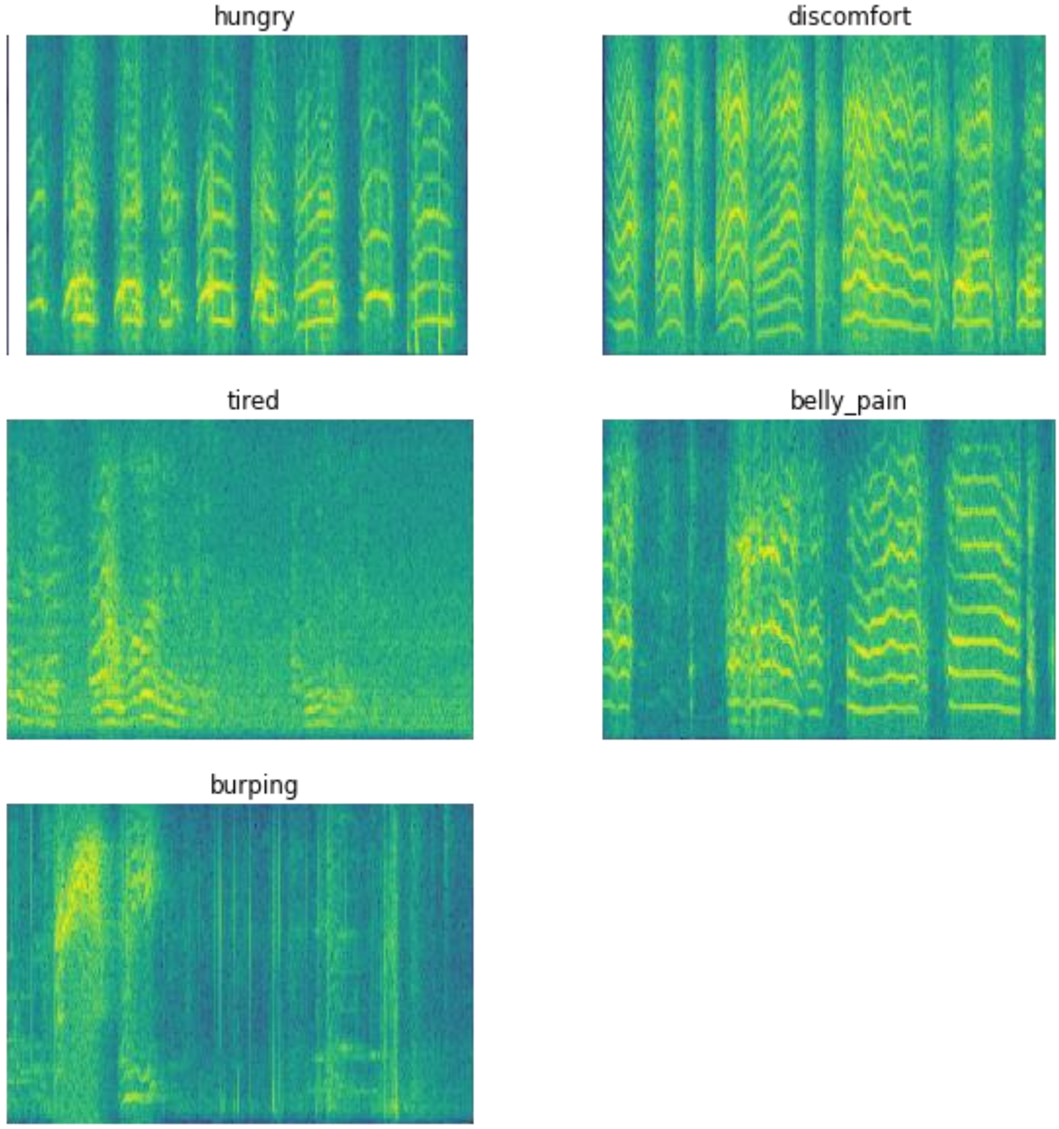


Figure 2. Example of Data Spectrograms

There is then a convolutional neural network created which starts by resizing the data. From this point on there are differences in the code for the three different testing methods: one for a 2-layer neural network, the next for a support vector machine, and finally a long short-term memory module. For the 2-layer neural network, the images of the spectrograms are first resized, then each pixel is normalized through a normal layer. Next, there are two convolutional layers. They are used because they can extract patterns and features from the pixels. Convolutional layers use a filter or a set of weights that, when multiplied by the input, showcases the probability of that patch of pixels representing a feature the filter is trying to find. This filter scans the image before moving a certain number of pixels and scanning once again. That movement is called a "stride". The output of one of these layers is a smaller image with more showcased patterns.

Both of the above-described layers, as well as other convolutional layers, use the Rectified Linear Unit (Relu). Deep learning modules that use a gradient algorithm tend to get trapped at a local minimum, in order to avoid this, the Relu function is used which speeds up the convergence learning of the module [9]. Following those is a pooling layer, which reduces the image even further by taking the largest value in an area and passing only that value forward to the next level. Later in this sequential model, two dense layers are used. Dense layers are layers in which every neuron in the previous layer is connected and sends a signal to each neuron in the dense layer.

The SVM (Support Vector Machine) model starts similarly with a resizing and normal layer. There are then two convolutional layers and a pooling layer. Afterward, Random Fourier Features using the gaussian radial basis function distributing parameters maps out, from the input layer's dimensions to lower dimensions to create a randomized feature space based on the approximate shift-invariant kernels. The SVM model is finished with a final dense layer.

The last model is an LSTM (Long Short-Term Memory) model also starting with resizing and a normal layer. The images are then reshaped, and a convolutional LSTM layer is applied. When doing this the image of a spectrogram is converted into a sequence of parts of images that are then passed to a convolutional layer before being passed to an LSTM which spots trends in this sequence and predicts the label. The learning rate of these models can be seen represented in Figure 3. This is used to stop the model training at a global minimum rather than a local one.

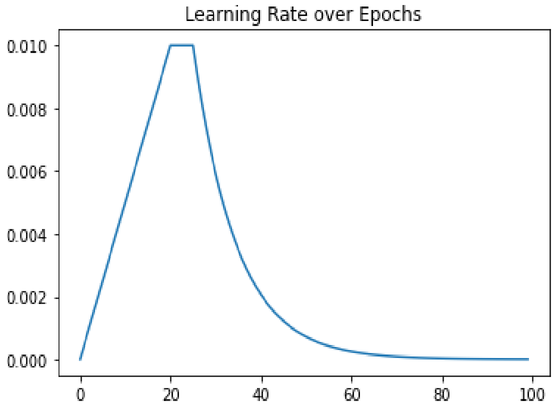


Figure 3. Learning Rate of Models over Epochs

## 3. Experiment

To determine which model was best in this situation, the accuracies of each were compared as shown in the below figure over a total of 100 epochs. For proper results, the three programs were all executed on Google Colab with the hardware accelerator being a GPU. It can be seen in Figure 4 that the SVM analyzed the data with higher accuracy of 86% when compared to the 2-layer neural network and LSTM with 81%.

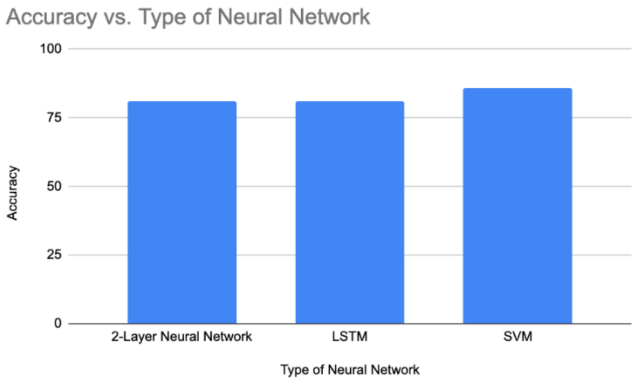


Figure 4. Accuracies of the Three Models for the Testing Set

For each of the three models, the sparse categorical cross-entropy loss function from logits is used when there are more than two label types. The SVM reported much lower levels of loss with just 1.0053 when compared to the other two models as the 2-layer neural network reported 3.53035 and the LSTM, the greatest of the three, with 3.8031. These results can be seen in Figures 5, 6, and 7. The accuracies of all three on the validation sets can also be seen in the graph with the SVM having an accuracy of 77.5%, the 2-Layer Neural Network with 75%, and the LSTM model with 72.5%.

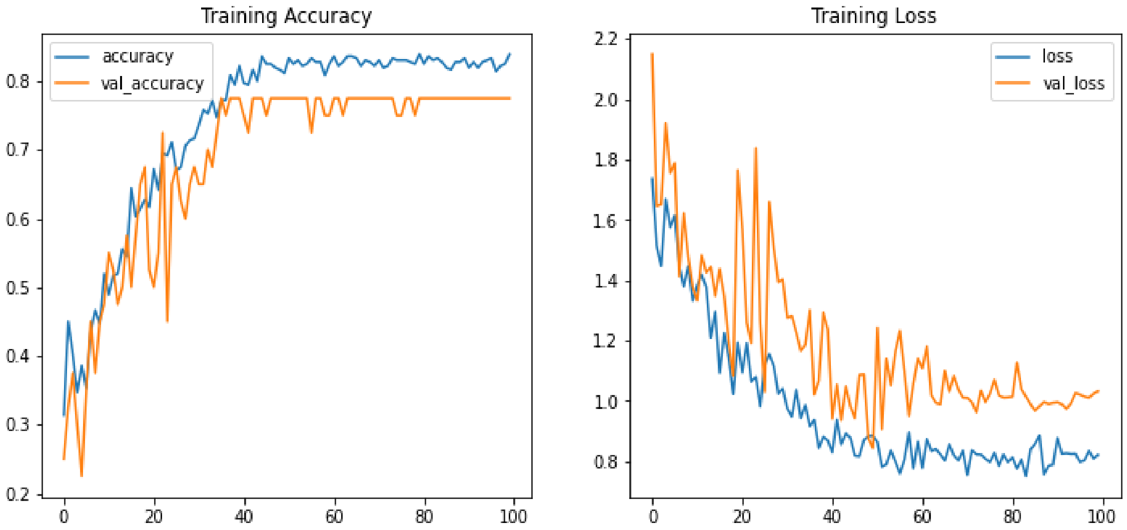


Figure 5. Accuracy and Loss Over Epochs of the SVM for Training and Validation Sets

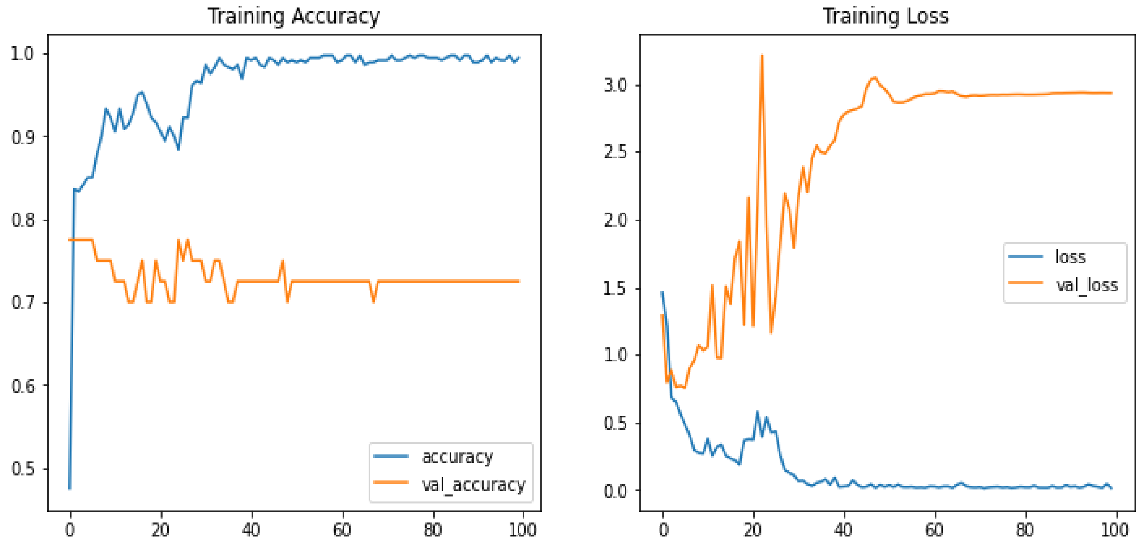


Figure 6. Accuracy and Loss Over Epochs of the 2-Layer Neural Network for Training and Validation Sets

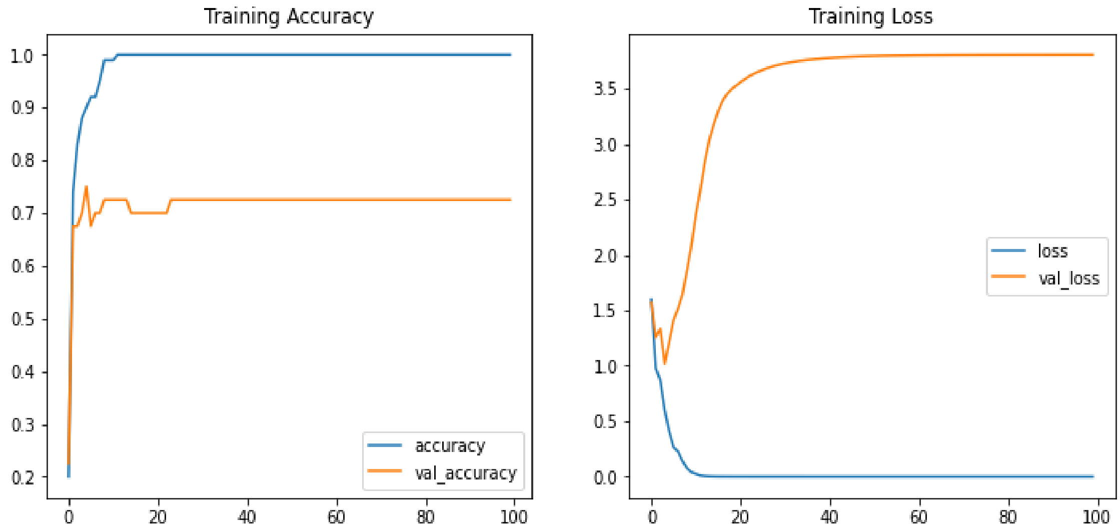


Figure 7. Accuracy and Loss Over Epochs of the LSTM for Training and Validation Sets

## 3.1. Analysis

Through the following analysis, it can be concluded the model most fit for determining the cause of a baby's cries is the SVM, which supports the highest accuracy without signs of severe overfitting. The relatively high accuracy of the SVM can be attributed to the high number of parameters as well as the small size of data. SVM’s can outperform neural networks when the data size is low, so this is one possible reason for the neural networks' relatively low accuracy. SVM’s perform at such high levels in small data sets because they create a hyperplane separating the observations they make [9]. This is useful because the SVM will always find a solution when classifying by increasing the dimensionality of its hyperplane to a level where there is a solution [10]. Another reason for the higher accuracy achieved when using the SVM is the number of layers. Given that if the number of parameters in the 2-layer neural network was increased to match that of the SVM, the 2-layer neural network should perform at a higher level of complexity, offering a much-improved accuracy.

As for the LSTM, there is another reason why it performed poorly. LSTM specializes in predicting the next picture in a sequence of pictures, or in a video, so it is out of its depth trying to spot trends and assign those trends to labels. To combat this each image of a spectrogram was split into parts of the whole and used as parts in a sequence in order to imitate a video, but to no benefit beyond what the other models would provide. Another possible hindrance to the LSTM model could be the lack of a second convolutional layer. Both of the other two models contained two convolutional layers compared to the LSTM model's one. Without this second layer, pulling and analyzing patterns and trends is made more difficult, potentially leading to a lower accuracy once the model is required to identify trends. While the validation accuracy is lower than that of the SVM's, the testing accuracy does reach perfection. This, along with the loss function graph, indicates severe overfitting. One explanation for this is that LSTM models have the ability to hold memory, so overtraining on a small size of data will lead to the model memorizing the data rather than learning the trends that each piece demonstrates.

Overfitting can also be seen in the case of the neural network. This can be seen in Figure 6 with the loss being much higher in the validation set than in the training set. While, as stated earlier, an increased number of parameters can be the catalyst that allows for increased accuracy and complexity, a high number of trainable parameters can lead to overfitting, which is possibly what is being seen here. The number of trainable parameters possessed by the 2-layer neural network is over ten times larger than either of the other two modules. Besides the SVM's tendency to thrive in small data size experiments, the number of trainable parameters could be a reason for not overfitting, in contrast to the 2-layer neural network which contains twenty times the number of trainable parameters.

## 4. RELATED WORKS

Other methods have been used in combination with a similar premise. One such example is by a group from Yunlin University who used a combination of convolutional neural networks to determine if the spectrogram fed to it was a baby's cry and then classify that cry's cause as one of four reasons [11]. Another group Universitas Indonesia paired convolutional neural networks with recurrent neural networks, allowing a more streamlined process in which the recurrent neural network learns off of the features extracted by the convolutional neural network [12]. In the same vein, a group from Georgia State University focused on convolutional neural networks as well, trying to improve upon them by using a multi-stage convolutional neural network with a hybrid feature set and prior knowledge [13].

In opposition to these ideas, a group from Koç University used a capsule network in direct comparison to regular convolutional neural networks and fed the audio signals from spectrograms created by their audio file data into the capsule network [14].

The first three models all use convolutional neural networks as a central part of their model, while the second and third ones try to improve upon the normal convolutional network by introducing new features, but still in hopes that a convolutional neural network is best suited for this task. The fourth group takes a different approach. Instead, this group compares a capsule network to convolutional neural networks in support of their model.

## 5. CONCLUSION

In order to solve the problem of determining the cause behind a baby's cry, first, a dataset with labeled reasons for why babies were crying was found. From here each file was assigned a label before being transformed into a waveform so that it could be seen and interpreted by programs. Waveform trends were also barely noticeable to the human eye inside of one label. In order to make these trends more obvious, each file was again transformed, now as a spectrogram. The spectrogram adds another layer that can be analyzed, making the trends more obvious to the programs. From here, three different programs were trained and tested on this data: a support vector machine, a 2-layer neural network, and a long short-term memory model (all of which incorporated at least one convolutional layer). After testing was complete, due to the highest accuracy and lowest loss, it was determined that for this experiment, the SVM was best suited to label the causes for babies to cry.

### 5.1. Current Limitations

There were some shortcomings to this experiment. The most glaring is the data size which disproportionately negatively affected the 2-layer neural network and long short-term model. Another place for improvement would be the number of reasons for which the baby is crying because they are currently limited to five. In practicality, there are a multitude of reasons that cause a baby to cry and limiting that number to five while training will lead to a much lower accuracy when being tested in real-life situations.

### 5.2. Future Works

In the future in order to improve upon this research, it would be beneficial to have access to a much larger data size to fairly test each module and accurately determine which model is optimal for this task. Finally, adding on as many causes to the cries is another step in improving this paper, which will significantly improve the accuracy when the testing is performed on actual babies.

## REFERENCES

[1] Foote, Keith D. "A Brief History of Natural Language Processing (Nlp)." DATAVERSITY, 17 June 2019, www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/.

[2] Deng, Li. *Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey*. 2012, www-edlab.cs.umass.edu/cs697l/readings/Three%20Classes%20of%20Deep%20Learning%20Architectures.pdf.

[3] Rajput, Adil. "Natural Language Processing, Sentiment Analysis, and Clinical Analytics." Innovation in Health Informatics, Academic Press, 15 Nov. 2019, www.sciencedirect.com/science/article/pii/B9780128190432000034.

[4] Kurth, Elisabeth, et al. "Crying Babies, Tired Mothers: What Do We Know? A Systematic Review." Midwifery, Churchill Livingstone, 20 Sept. 2009, www.sciencedirect.com/science/article/abs/pii/S0266613809000692.

[5] K. Sharma, C. Gupta and S. Gupta, "Infant Weeping Calls Decoder using Statistical Feature Extraction and Gaussian Mixture Models," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944527.

[6] Reynolds, Douglas. *Gaussian Mixture Models*. 2009, leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/refs/GMM_Tutorial_Reynolds.pdf.

[7] Gveres. “Gveres/Donateacry-Corpus: An Infant Cry Audio Corpus That's Being Built through the Donate-a-Cry Campaign - See Http://Donateacry.com.” GitHub, github.com/gveres/donateacry-corpus.

[8] Chaudhary, Kartik. “Understanding Audio Data, Fourier TRANSFORM, FFT, Spectrogram and Speech Recognition.” *Medium*, Towards Data Science, 18 Jan. 2020, towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520.

[9] K. Hara, D. Saito and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1-8, doi: 10.1109/IJCNN.2015.7280578.

[10] Luca, Gabriele De. “SVM vs Neural Network.” Baeldung on Computer Science, 9 Sept. 2020, www.baeldung.com/cs/svm-vs-neural-network.

[11] Chang CY., Tsai LY. (2019) A CNN-Based Method for Infant Cry Detection and Recognition. In: Barolli L., Takizawa M., Xhafa F., Enokido T. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing, vol 927. Springer, Cham. https://doi.org/10.1007/978-3-030-15035-8_76

[12] Maghfira1, Tusty Nadia, et al. “Iopscience.” *Journal of Physics: Conference Series*, IOP Publishing, 1 Apr. 2020, iopscience.iop.org/article/10.1088/1742-6596/1528/1/012019.

[13] Ji C., Basodi S., Xiao X., Pan Y. (2020) Infant Sound Classification on Multi-stage CNNs with Hybrid Features and Prior Knowledge. In: Xu R., De W., Zhong W., Tian L., Bai Y., Zhang LJ. (eds) Artificial Intelligence and Mobile Services – AIMS 2020. AIMS 2020. Lecture Notes in Computer Science, vol 12401. Springer, Cham. https://doi.org/10.1007/978-3-030-59605-7_

[14] Tug˘tekinTuran, and ErzinEngin. *Monitoring Infant’s Emotional Cry in Domestic Environments Using the Capsule Network Architecture*. Sept. 2018

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# BUILDING A SMART MIRROR FOR THE PURPOSES OF INCREASED PRODUCTIVITY AND BETTER MENTAL HEALTH, COMPLETE WITH AN APP

Jonathan Liu[1] and Yu Sun[2]

[1]Arcadia High School, Arcadia, CA, 91007
[2]California State Polytechnic University, Pomona, CA, 91768

### *ABSTRACT*

*Oftentimes, people find themselves staring in the mirror mindlessly while brushing their teeth or putting on clothes. This time, which may seem unnoticeable at first, can accumulate to a significant amount of time wasted when looked at over the duration of a year, and can easily be repurposed to better suit one's goals. In this paper, we describe the construction and implementation of the Smart Mirror, an intelligent mirror that boasts several features in order to improve an individual's daily productivity. As the name suggests, it is a mirror, and so will not take anything away from the user when he or she is performing their daily teeth brushing. It also hosts facial recognition, and can recognize one's emotions from one glimpse through the camera. The Mirror also comes with an app that is available on the Google Play Store, which helps input tasks and daily reminders that can be viewed on the Smart Mirror UI.*

### *KEYWORDS*

*Thunkable, Google Firebase, Android, Raspberry Pi*

## 1. INTRODUCTION

Especially in the Covid-19 pandemic, but even in general, productivity has decreased for many, including students [1]. This can mainly be attributed to the heightened amount of time spent in front of a screen (whether it be for classes or business meetings) while at home. These computers, while they can be used to do work, also have the ability to play video games or watch forms of entertainment such as YouTube and Netflix. The distractions on laptops are immense, and contribute to decreased productivity because more students are drawn in by the prospect of gaming and less by the possibility of completing homework assignments. After all, which is more fun, video games or analyzing a 17th-century speech? Most students would answer with the former.

In addition, a small but noticeable amount of time is spent in front of the mirror. This time is not utilized to its maximum potential when someone just stares at them self while brushing their teeth. Other actions may be performed, but with the risk of distracting from the original purpose of brushing teeth. This thus presents individuals with a question: how does one increase their efficiency on a day-by-day basis, without being distracted in major ways?

The Smart Mirror aims to solve this issue of decreased productivity. Working as essentially a scaled-up version of a phone interface without apps and just a to-do list, its clean UI allows for

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 61-71, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111406

easy interpretation and usage for people seeking to improve and reach their daily goals. The Android app that accompanies the Mirror further simplifies the purpose of the gadget by providing the local time and other crucial features. The cherry on top is that it is a mirror, so it doesn’t distract from everyday tasks such as going out with friends, buying groceries, or driving to work or school. Thus, it is not meant to replace the average Google Calendar or Todoist (a to-do list application), but to provide a simple checklist and reminder of tasks that need to be done at the start and end of each day.

There have been very primitive prototypes of a “Magic Mirror” [2] in the past, such as one built by MIT [3] and one by Michael Teeuw [4], but they do not show much besides the basic clock, weather, and news. The user interface is also very messy and clunky, having the news on the left side and the clock directly in the middle. On screens with less than 13 inches of diagonal length, the Magic Mirror is no longer a mirror, but a mash of numbers for the clock and date and words from the news. It is neither aesthetically pleasing nor does it provide effective efficiency for users, as it completely eliminates the original purpose of the device - to still function as a mirror.

A number of social media users and YouTubers have posted videos on building a “Smart Mirror” [5], but they use the same Magic Mirror prototype code mentioned above. No videos seem to have any creators to add their own code to make it more functional and purposeful, and overall, more unique. In addition, in all these videos, the screens and glass used were significantly large, most over 24 inches diagonally. This allows for a better, more aesthetically pleasing interface on larger screens, but does not resolve the fact that the Magic Mirror's UI won’t work well on smaller displays.

Our device expands on the “Magic Mirror” [2] mentioned previously, and includes a Raspberry Pi, facial recognition, a to-do list, and Firebase connectivity to the list of features. The input data (for either the to-do list or reminders) gathered from the Android companion app is sent to a Firebase cloud database, which in turn sends these inputs to the Mirror itself. The tasks will then appear in the center-bottom of the Smart Mirror, making for easy visual access while also not distracting from the primary purpose of a regular mirror (to look at oneself). The Mirror also boasts a Raspberry Pi camera, which is used to both detect facial emotions and take a picture of the user if prompted to. Facial recognition is then used to output whatever emotion the algorithm thinks the user is emitting, whether it be joy or regret.

The app that accompanies the Smart Mirror is downloadable in the Google Play Store under the name “Smart Mirror Companion” [6]. It’s a simple yet effective app, as it allows individuals to input whatever tasks or reminders they need to complete for the day. And, as mentioned before, this data is sent to Google’s Firebase database [7], which the Mirror represents as a section in the center-bottom. Within the app, people can also check off tasks or reminders using the easy switch button. The reminders switch timestamps the input (and users can also edit this to fit their schedule), so people can see when their reminders need to be completed by.

From the list of features described above, it is clear that the Smart Mirror further diversifies and outperforms the skill set of Teeuw’s Magic Mirror.

In two experiments described more in detail in Section 4, we demonstrate how the Smart Mirror does, in fact, improve upon daily productivity in the long run. First, we take a look at a six-week period during which users compare the before-and-after of using the Smart Mirror. We analyze their satisfaction with it, and if they found their efficiency increased or decreased. Second, we break down a user who used the Mirror for one month after not using any productivity tools. We analyze her task memory (if she remembered to do tasks better with the Mirror versus not), and also show a graph displaying the results.

The rest of the paper is structured as follows: Section 2 details the related works which served as the basis of this project; Section 3 focuses on the challenges encountered in the study, and how they were subsequently overcome using preexisting and new methods; Section 4 describes the solutions to these challenges in greater, more intricate detail; Section 5 denotes some experiments we conducted; and Section 6, the last section, will provide a conclusion and offer insight into future work that be done.

## 2. RELATED WORK

The first, most obvious related piece of work would be the Magic Mirror [2, 3, 4]. It served as the basis of the Smart Mirror, and was very friendly in terms of its accessibility and knowledge base. The Magic Mirror documentation already provided the most simple necessities, including time, calendar, and a news feed. But, as pointed out above, it wasn't very user-friendly and certainly was not a mirror for small screens. The Smart Mirror solved this problem using text adjustment, and also includes more features (such as the to-do list and taskbar). It can clearly be seen that the Smart Mirror outclasses the Magic Mirror due to its vast array of features.

Other notable versions of the "Magic Mirror" mostly consist of YouTubers doing DIY projects [5]. Their mirrors, however, utilize the base that is the Magic Mirror mentioned as the first related work, and don't boast any creative or unique aspects and features like the Smart Mirror does.

The YouTubers' DIY videos mostly follow the same script: introduce the concept of a Magic Mirror, download the Magic Mirror documentation found on the Magic Mirror website, purchase and connect the Raspberry Pi, and finally, build the mirror itself using wood, one-sided glass, and glossy finishes [5]. No additional "personality" features were added, so the videos became repetitive and lost their appeal.

## 3. CHALLENGES

In order to build the Smart Mirror, an intelligent mirror that boasts several features to improve an individual's daily productivity, a few challenges have been identified as follows.

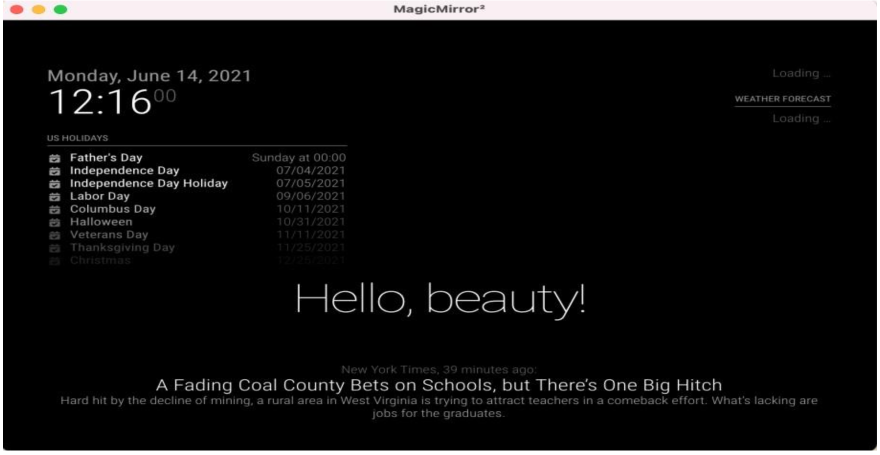
### 3.1. Challenge 1: Choosing a Method of Displaying the To-Do List

The first challenge was choosing a method of displaying the to-do list, one of the unique features we decided to implement into the Smart Mirror. After researching several options, we decided on Google's Firebase. It proved to be the most reliable and consistent in terms of displaying list items on the Mirror screen, and was also the simplest to use. However, even with this decision, there was still the issue of connecting to the mirror's program and actually displaying it in the user interface. This meant that our solution needed to contain two steps: one to include a reference to Firebase, and the other to take inputs and data from the database and display it on the screen.

To do this, we implemented Node.js's [12] request function, which asks for an https website link. The parameter was replaced with our Realtime Database link from the Firebase console. This then presented another problem: how could we load the list from the Firebase server, and render it in the Smart Mirror module?

### 3.2. Challenge 2: Persistent UI and Aesthetic Issues Requiring Code

The second challenge after implementing Firebase presented itself when we tried to run the program. The primary device at the time was a MacBook Air, and when our edited Magic Mirror ran on Electron [8], the interface was found to be somewhat clunky and could not be personalized (i.e., specific panels could not be added or deleted; see Figure 1). The obvious solution was to switch to a larger screen, but the original problem still persisted (users with smaller screens would lose the "mirror" aspect of the device and instead just view text). We decided to utilize proportions to attempt to fix this issue, and this meant quite a few lines of code to fix the text adjustment and tile sizes.



MagicMirror$^2$

Monday, June 14, 2021

12:16$^{00}$

Loading ...

WEATHER FORECAST

Loading ...

US HOLIDAYS

| | |
|---|---|
| Father's Day | Sunday at 00:00 |
| Independence Day | 07/04/2021 |
| Independence Day Holiday | 07/05/2021 |
| Labor Day | 09/06/2021 |
| Columbus Day | 10/11/2021 |
| Halloween | 10/31/2021 |
| Veterans Day | 11/11/2021 |
| Thanksgiving Day | 11/25/2021 |
| Christmas | 12/25/2021 |

Hello, beauty!

New York Times, 39 minutes ago:

A Fading Coal County Bets on Schools, but There's One Big Hitch

Hard hit by the decline of mining, a rural area in West Virginia is trying to attract teachers in a comeback effort. What's lacking are jobs for the graduates.

Figure 1. Magic Mirror display

The original Magic Mirror display is shown in Figure 1. As can be seen, there is no longer the "mirror" aspect with a small display.

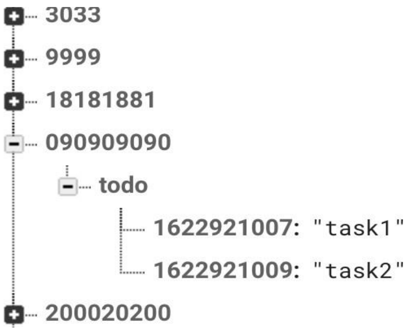
### 3.3. Challenge 3: Logistic Challenges of Building the Mirror

The third challenge was building the mirror itself. This proved to be harder than initially thought, as it required detailed planning and organization of where each component was to go. Parts such as the Raspberry Pi and its accompanying camera, along with its wires, needed to be placed in spots where it wouldn't make the mirror seem too bulky or thick, for example. One-sided glass was also in limited supply because of the pandemic. In addition, the Raspberry Pi was somewhat loud when the Smart Mirror program ran, so we had to include additional cooling features in order to lower the sound (because no one wants a loud fan whirring behind a mirror).

## 4. Our Proposed Solution

First, we implemented the code found in our research for a Smart Mirror, which can be found on the Magic Mirror website. This provided us with all the basic necessities, such as the clock and news panels (the portions of the mirror that were already visible when the program was run). After a quick scan of the folders included within the Magic Mirror program, we deduced that the "modules" folder would be our main focus, as it was the folder that contained all the "customizing" that needed to be done (including the camera, to-do list implementation, etc.)

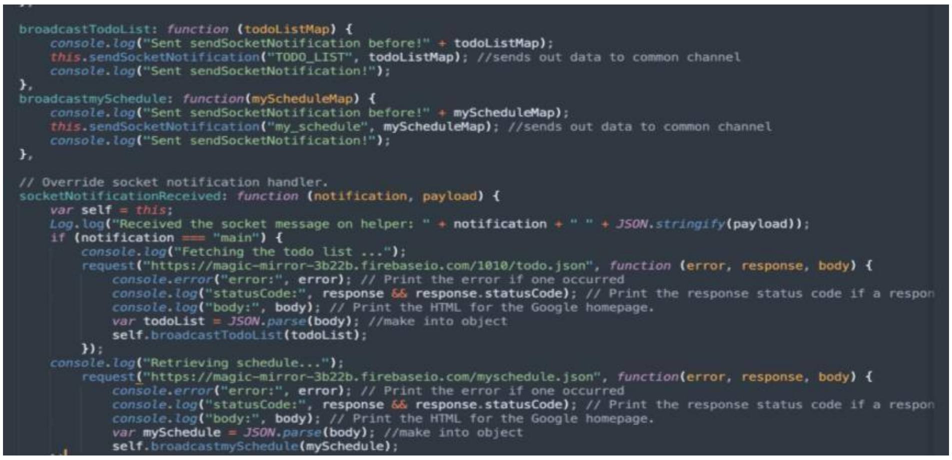
Second, Google Firebase needed to be integrated into the program. We chose Firebase because it was the simplest and relatively easy to understand. Backend data could also be easily viewed by a few clicks, so not much effort is required (see Figure 2).



```
3033
9999
18181881
090909090
    todo
        1622921007: "task1"
        1622921009: "task2"
200020200
```

Figure 2. A screenshot of the Realtime Database in Firebase. The numbers represent test device ID’s, and the text under “todo” represent the inputs entered by the user.

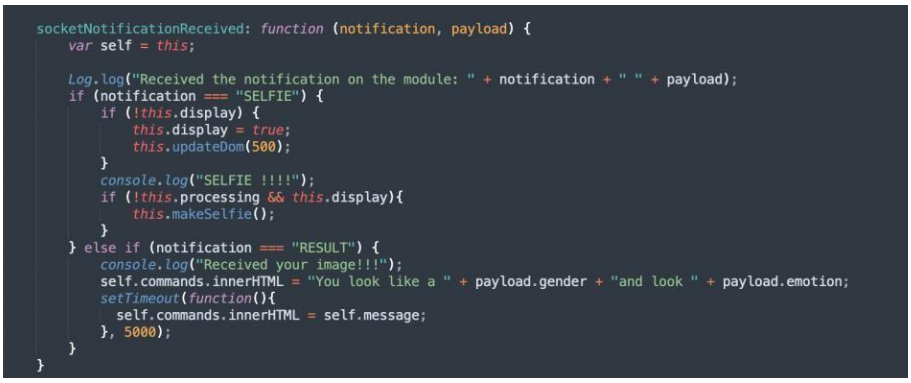
We implemented this and the Firebase into our code, as shown in Figure 3. The broadcastTodoList() function displays the To-do List found in the Firebase Realtime Database.



```
broadcastTodoList: function (todoListMap) {
    console.log("Sent sendSocketNotification before!" + todoListMap);
    this.sendSocketNotification("TODO_LIST", todoListMap); //sends out data to common channel
    console.log("Sent sendSocketNotification!");
},
broadcastmySchedule: function(myScheduleMap) {
    console.log("Sent sendSocketNotification before!" + myScheduleMap);
    this.sendSocketNotification("my_schedule", myScheduleMap); //sends out data to common channel
    console.log("Sent sendSocketNotification!");
},

// Override socket notification handler.
socketNotificationReceived: function (notification, payload) {
    var self = this;
    Log.log("Received the socket message on helper: " + notification + " " + JSON.stringify(payload));
    if (notification === "main") {
        console.log("Fetching the todo list ...");
        request("https://magic-mirror-3b22b.firebaseio.com/1010/todo.json", function (error, response, body) {
            console.error("error:", error); // Print the error if one occurred
            console.log("statusCode:", response && response.statusCode); // Print the response status code if a respon
            console.log("body:", body); // Print the HTML for the Google homepage.
            var todoList = JSON.parse(body); //make into object
            self.broadcastTodoList(todoList);
        });
    console.log("Retrieving schedule...");
        request("https://magic-mirror-3b22b.firebaseio.com/myschedule.json", function(error, response, body) {
            console.error("error:", error); // Print the error if one occurred
            console.log("statusCode:", response && response.statusCode); // Print the response status code if a respon
            console.log("body:", body); // Print the HTML for the Google homepage.
            var mySchedule = JSON.parse(body); //make into object
            self.broadcastmySchedule(mySchedule);
```

Figure 3. A snippet of the added code.

The next step was to include the camera feature, as it was an integral part of the creative and unique aspect of the Smart Mirror. To do this, we considered whether to create our own facial recognition program, or just find a self-sustaining program on GitHub [9] and paste that into the mirror’s program. After some research, we decided on the latter because it would save a significant amount of time that we could use to improve other aspects of the Smart Mirror. Some detailed facial recognition examples were found on GitHub, but we settled on a seemingly complicated, yet simple program because it boasted the ability to detect emotion and print it onto the screen [11]. It also wasn’t large in file size, which was important due to the Raspberry Pi and its loud fans. From there, it fit right in with the “modules” folder in the Smart Mirror, and so allowed for easy access if any revisiting of the code was needed (see Figure 4).

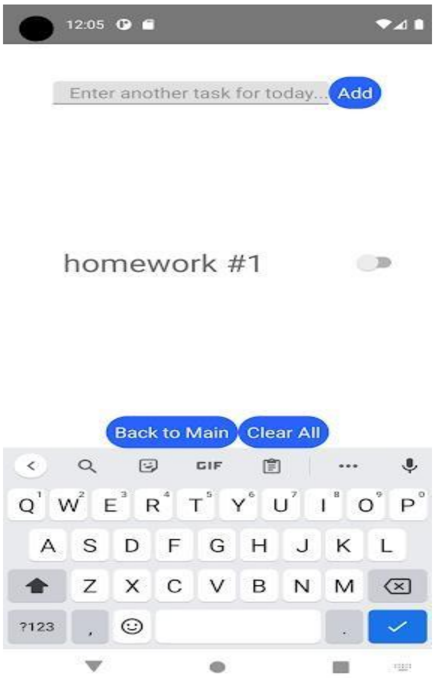


```
socketNotificationReceived: function (notification, payload) {
    var self = this;

    Log.log("Received the notification on the module: " + notification + " " + payload);
    if (notification === "SELFIE") {
        if (!this.display) {
            this.display = true;
            this.updateDom(500);
        }
        console.log("SELFIE !!!!");
        if (!this.processing && this.display){
            this.makeSelfie();
        }
    } else if (notification === "RESULT") {
        console.log("Received your image!!!");
        self.commands.innerHTML = "You look like a " + payload.gender + "and look " + payload.emotion;
        setTimeout(function(){
          self.commands.innerHTML = self.message;
        }, 5000);
    }
}
```

Figure 4. A portion of the camera program that prints out the user's gender and / or facial emotions after taking a selfie.

Next was the usage of a Raspberry Pi to be able to run the program. Raspberry Pi was recommended by the Magic Mirror creators, and it's also small enough to not be any large hindrance in terms of size or width of the actual mirror itself. A Raspberry Pi Camera is also used since it pairs with the Raspberry Pi and does not require extra steps to connect to the Pi. We followed the steps on the Raspberry Pi website [10] to build it. The three heatsinks that were included in the Raspberry Pi package were placed onto the chips so that the fan didn't have to work extremely hard (and thus produce overly loud sound) to cool the device.

While the Smart Mirror is not touchscreen, it can be controlled via a mouse if the user needs to. However, much of the actions can be done through the Smart Mirror Companion App, which allows users to input their daily tasks and reminders. These list items will then be sent to the Mirror and will show in the user interface. As such, there is really no need for a mouse.



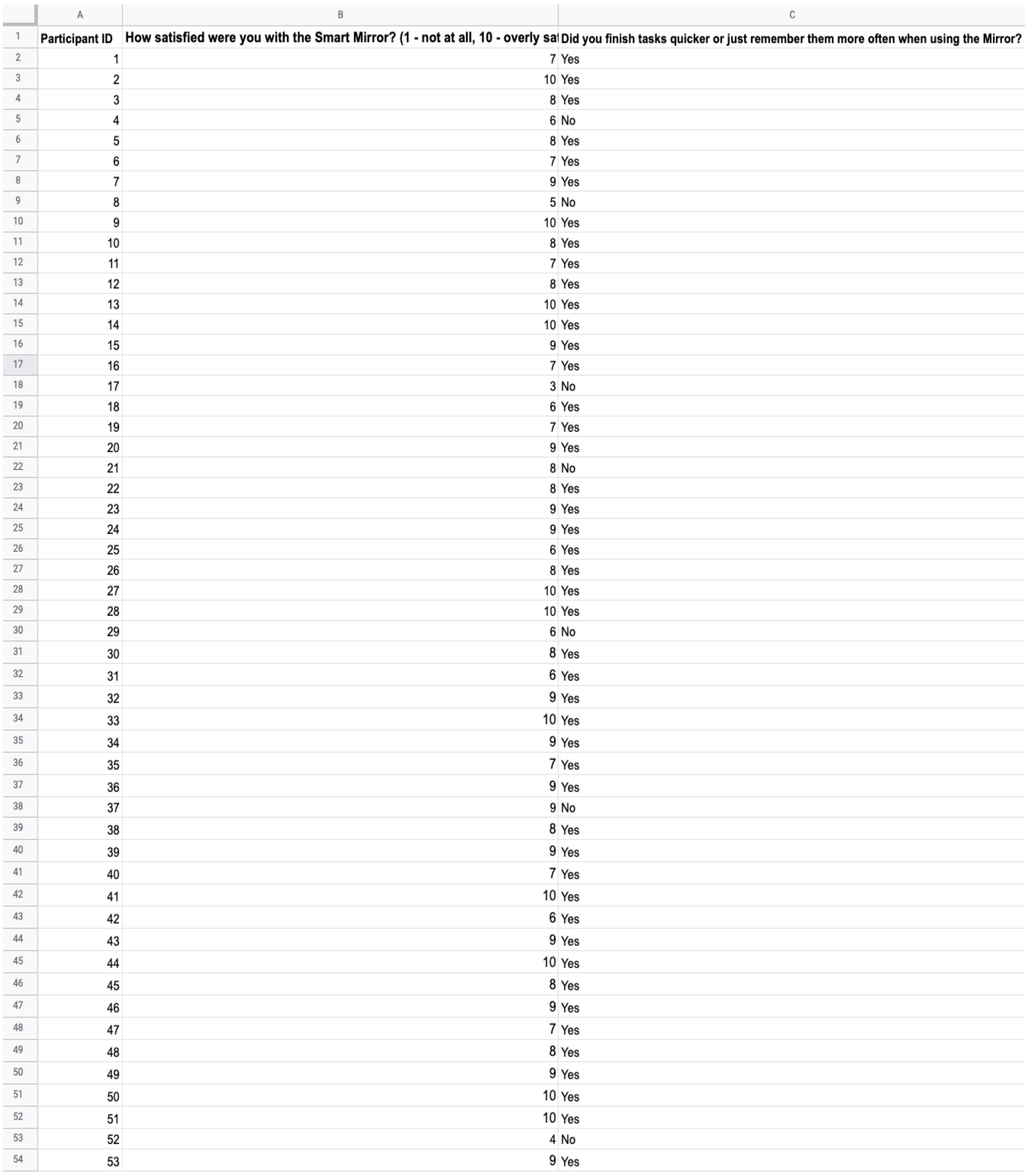
homework #1

Figure 5. Screenshot of the App, which can be found on the Google Play Store; "Enter text" box allows users to type in their tasks, which will be displayed in the Mirror interface.

# 5. Experiments

## 5.1. Experiment 1

To provide a quantifiable experiment to prove improvement with the Smart Mirror, we first asked around two hundred people to see if they would like to participate. About 53 responded. In this experiment, we wanted to see if the Mirror really lived up to the above claims of increased efficiency. Over a three-week period, participants were surveyed weekly on how efficiently they were able to complete or simply just remember tasks. Then, in another three-week period, the same people used the Smart Mirror and continued to perform their daily routines. After the entire experiment was over, another survey was sent out, asking the people if they noticed that their memory improved or were able to complete tasks earlier on in the day. 46 people, or about 87%, individuals out of the 53 noted that they were able to both remember and finish tasks off more easily when using the Smart Mirror (see Figure 6).



| | A | B | C |
|---|---|---|---|
| 1 | Participant ID | How satisfied were you with the Smart Mirror? (1 - not at all, 10 - overly sat | Did you finish tasks quicker or just remember them more often when using the Mirror? |
| 2 | 1 | 7 | Yes |
| 3 | 2 | 10 | Yes |
| 4 | 3 | 8 | Yes |
| 5 | 4 | 6 | No |
| 6 | 5 | 8 | Yes |
| 7 | 6 | 7 | Yes |
| 8 | 7 | 9 | Yes |
| 9 | 8 | 5 | No |
| 10 | 9 | 10 | Yes |
| 11 | 10 | 8 | Yes |
| 12 | 11 | 7 | Yes |
| 13 | 12 | 8 | Yes |
| 14 | 13 | 10 | Yes |
| 15 | 14 | 10 | Yes |
| 16 | 15 | 9 | Yes |
| 17 | 16 | 7 | Yes |
| 18 | 17 | 3 | No |
| 19 | 18 | 6 | Yes |
| 20 | 19 | 7 | Yes |
| 21 | 20 | 9 | Yes |
| 22 | 21 | 8 | No |
| 23 | 22 | 8 | Yes |
| 24 | 23 | 9 | Yes |
| 25 | 24 | 9 | Yes |
| 26 | 25 | 6 | Yes |
| 27 | 26 | 8 | Yes |
| 28 | 27 | 10 | Yes |
| 29 | 28 | 10 | Yes |
| 30 | 29 | 6 | No |
| 31 | 30 | 8 | Yes |
| 32 | 31 | 6 | Yes |
| 33 | 32 | 9 | Yes |
| 34 | 33 | 10 | Yes |
| 35 | 34 | 9 | Yes |
| 36 | 35 | 7 | Yes |
| 37 | 36 | 9 | Yes |
| 38 | 37 | 9 | No |
| 39 | 38 | 8 | Yes |
| 40 | 39 | 9 | Yes |
| 41 | 40 | 7 | Yes |
| 42 | 41 | 10 | Yes |
| 43 | 42 | 6 | Yes |
| 44 | 43 | 9 | Yes |
| 45 | 44 | 10 | Yes |
| 46 | 45 | 8 | Yes |
| 47 | 46 | 9 | Yes |
| 48 | 47 | 7 | Yes |
| 49 | 48 | 8 | Yes |
| 50 | 49 | 9 | Yes |
| 51 | 50 | 10 | Yes |
| 52 | 51 | 10 | Yes |
| 53 | 52 | 4 | No |
| 54 | 53 | 9 | Yes |

Figure 6. Satisfaction with the Smart Mirror ranked on a scale of 1-10 for 50(+) participants.

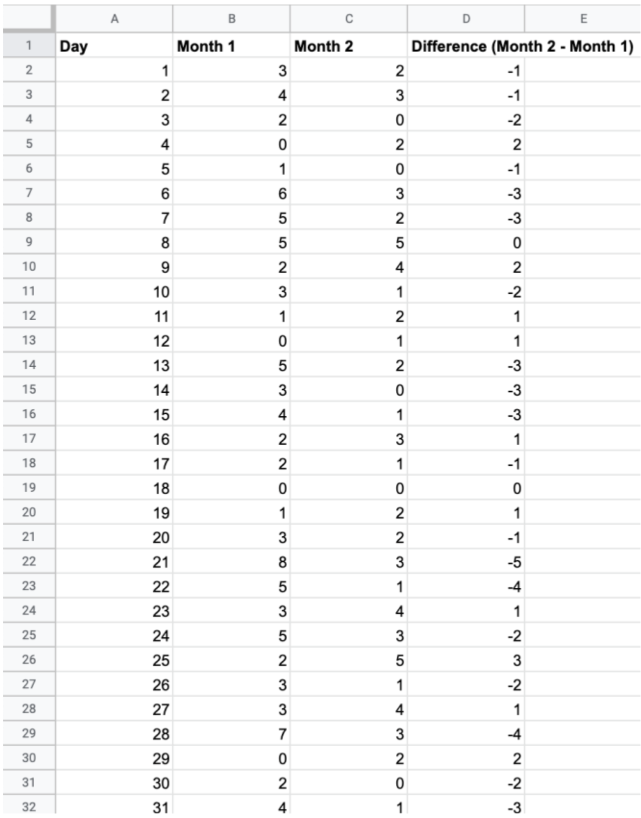
The first question was merely a survey question that allowed us to access feedback based on 53 users. Many scored the Mirror above 5 (a good sign, of course), with some finding it less than ideal and not living up to expectations. The average value of these participants' scores was 8.07, a high score for such an early version of a product. In addition, in a separate "feedback" section, users told us what they would have liked to see after using the Smart Mirror. This will be discussed in the last paragraph of this subsection.

From the table data provided, it is quite clear that while many users found the Smart Mirror unsatisfactory, participants still were able to finish tasks quicker, or, more generally, remember them when using the Mirror. Only 7 people responded "no" to the second question, which is a small minority of the total participants. However, these seven individuals all did not rate the Mirror "badly" (i.e. scores of 5 or less); some gave the Mirror a relatively high score. It can be inferred that the Mirror itself was useful and interesting, but the "productivity" aspect was not as beneficial to them.

The fact that many users thought that the Mirror was unsatisfactory can most probably be attributed to the still-clunky user interface. While our version improves on the interface provided by the Magic Mirror (it introduces text adjustment, for example), it still does not allow people to hide or add additional panels. For instance, if one wanted to integrate their own calendar and sync it with Google Calendar, the Smart Mirror cannot do that. As such, it is understandable why many submissions put scores of 5 or lower as the answer to the first question ("How satisfied were you with the Smart Mirror? (1 - not at all, 10 - overly satisfied)"). This feature of adding/eliminating panels is part of our future work on the Smart Mirror (further information in Section 6).

## 5.2. Experiment 2

Our second experiment involved only one person, but lasted two months. The idea was to have one person record the number of tasks she forgot within a month (each day counting how many tasks she forgot, for 31 days). Then, in the second month, the subject used the Smart Mirror and continued to log the number of forgotten tasks. Each number was recorded in a Google Spreadsheet (see Figure 7).



|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Day | Month 1 | Month 2 | Difference (Month 2 - Month 1) |  |
| 2 | 1 | 3 | 2 | -1 |  |
| 3 | 2 | 4 | 3 | -1 |  |
| 4 | 3 | 2 | 0 | -2 |  |
| 5 | 4 | 0 | 2 | 2 |  |
| 6 | 5 | 1 | 0 | -1 |  |
| 7 | 6 | 6 | 3 | -3 |  |
| 8 | 7 | 5 | 2 | -3 |  |
| 9 | 8 | 5 | 5 | 0 |  |
| 10 | 9 | 2 | 4 | 2 |  |
| 11 | 10 | 3 | 1 | -2 |  |
| 12 | 11 | 1 | 2 | 1 |  |
| 13 | 12 | 0 | 1 | 1 |  |
| 14 | 13 | 5 | 2 | -3 |  |
| 15 | 14 | 3 | 0 | -3 |  |
| 16 | 15 | 4 | 1 | -3 |  |
| 17 | 16 | 2 | 3 | 1 |  |
| 18 | 17 | 2 | 1 | -1 |  |
| 19 | 18 | 0 | 0 | 0 |  |
| 20 | 19 | 1 | 2 | 1 |  |
| 21 | 20 | 3 | 2 | -1 |  |
| 22 | 21 | 8 | 3 | -5 |  |
| 23 | 22 | 5 | 1 | -4 |  |
| 24 | 23 | 3 | 4 | 1 |  |
| 25 | 24 | 5 | 3 | -2 |  |
| 26 | 25 | 2 | 5 | 3 |  |
| 27 | 26 | 3 | 1 | -2 |  |
| 28 | 27 | 3 | 4 | 1 |  |
| 29 | 28 | 7 | 3 | -4 |  |
| 30 | 29 | 0 | 2 | 2 |  |
| 31 | 30 | 2 | 0 | -2 |  |
| 32 | 31 | 4 | 1 | -3 |  |

Figure 7. Results of Experiment 2; forgotten tasks without Smart Mirror (month one) and with Smart Mirror (month two)

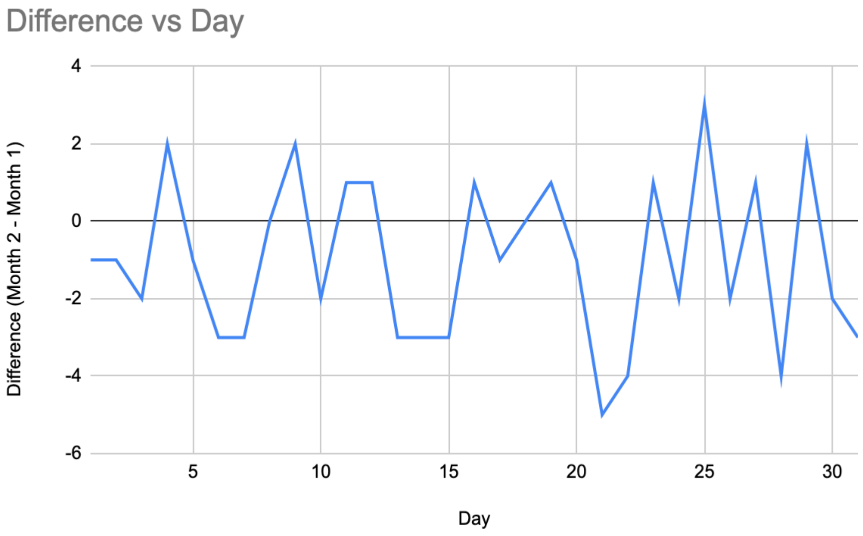


Figure 8. The graph of the data collected in Experiment 2.

As seen in the graph, much of the data points lie below the 0 line, meaning more tasks weren't forgotten when using the Smart Mirror. Thus, it can be claimed that the Smart Mirror improves productivity, as the person remembered to do more tasks when using the Mirror. We also provided a feedback form to the participant to see if there was anything we could improve on. Her only suggestion was for us to include more customizable options, as she didn't really need the Holidays panel. This is the common denominator in the previous experiment: people want the

Mirror to be customizable and to be able to add or remove certain parts of the Mirror. And, as mentioned, this will be discussed in Section 6, “Future Works.”

These two experiments were aimed to prove the idea that the Smart Mirror improves productivity. And, judging from the results of both experiments, it is quite clear that the Mirror does what we intended it to do. It not only allows people to finish tasks quicker (results of Experiment 1), it also just simply lets users remember them more easily (results of Experiment 2). Thus, it reaches our expectations, but just barely. There are still a number of more things to work on, such as allowing users to add or remove panels.

## 6. CONCLUSION AND FUTURE WORK

The Smart Mirror is a new and improved version of the Magic Mirror built by inventor Michael Teeuw [4]. Initially conceived to solve the problem of decreased productivity (due to the effects of the pandemic), it grew into a personal project where we wanted to customize the mirror and allow others besides ourselves to enjoy it. As such, it possesses more features, including a to-do task bar and calendar. It adds to the list of features found with the Magic Mirror (time, news feed, major holidays [15, 16]). In order to include our desired additional features, we needed to connect Google’s Firebase and design an additional app, which is called the “Smart Mirror Companion App” [6], which can be found on the Google Play Store. Automatic text adjustment was also added so that smaller screens could enjoy the Smart Mirror and not be hindered by large amounts of text.

The most significant limitation was the building of the mirror. It required a frame (wood or metal), a Raspberry Pi and accompanying camera, and one-sided glass. On top of that, it also required some building or craftsmanship skills. After all, a list of materials is only a list; there is still the issue of building the mirror itself, and making sure nothing goes wrong during the process. For people who are not skilled with their hands or have no experience in building things, this would be a tedious procedure.

In the near and distant futures, more thought will be directed towards optimizing the build quality of the mirror (such as the optimal placement of the Raspberry Pi) while still allowing unskilled persons to successfully build their own Smart Mirror. If the Smart Mirror were ever to be marketed and sold, this issue would definitely need to be solved beforehand so as to provide the most customer satisfaction.

In addition to these additions, more unique features can be added upon request. One such feature that we are already thinking about adding is connecting Google Calendar [13] or Todoist [14] to the Mirror. This would completely eliminate the need for our additional app, as it can just display events or tasks that users can input easily via their computer.

## AUTHORS’ NOTES

Although many reviewers have requested the GitHub for the Smart Mirror (and we respect their thoughts and inputs), we would like to keep it private in case the Mirror becomes a marketable product in the future.

## REFERENCES

[1] “Gendered Effects of Covid-19 on Scientific Productivity” - https://journals.sagepub.com/doi/full/10.1177/23780231211006977

[2] MagicMirror - https://magicmirror.builders/

[3] MIT Magic Mirror - https://courses.media.mit.edu/2016spring/mass65/2016/05/14/the-magic-mirror/

[4] Inventor of the MagicMirror, Michael Teeuw - https://michaelteeuw.nl/

[5] Example of YouTube “DIY Smart Mirror” - https://youtu.be/OYlloiaBINo

[6] “Smart Mirror Companion App”, Google Play - https://play.google.com/store/apps/details?id=com.gmail.jj2004liu.magicmirrorv1

[7] Google Firebase - https://firebase.google.com/

[8] Electron - https://www.electronjs.org/

[9] Facial Recognition Programs on Github - https://github.com.cnpmjs.org/topics/face-recognition

[10] Raspberry Pi Camera Usage - https://raspberrytips.com/install-camera-raspberry-pi/

[11] Github, MagicMirror Camera Module - https://github.com/alexyak/camera

[12] Node.js - https://nodejs.org/en/

[13] Google Calendar - https://calendar.google.com/

[14] Todoist, one of the most popular to-do checklist apps - https://todoist.com//

[15] Newsfeed - https://rss.nytimes.com/services/xml/rss/nyt/HomePage.xml

[16] Major Holidays - https://www.timeanddate.com/holidays/us/

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# ENSEMBLE CREATION USING FUZZY SIMILARITY MEASURES AND FEATURE SUBSET EVALUATORS

Valerie Cross and Michael Zmuda

Computer Science and Software Engineering,
Miami University, Oxford, OH, USA

### *ABSTRACT*

*Current machine learning research is addressing the problem that occurs when the data set includes numerous features but the number of training data is small. Microarray data, for example, typically has a very large number of features, the genes, as compared to the number of training data examples, the patients. An important research problem is to develop techniques to effectively reduce the number of features by selecting the best set of features for use in a machine learning process, referred to as the feature selection problem. Another means of addressing high dimensional data is the use of an ensemble of base classifiers. Ensembles have been shown to improve the predictive performance of a single model by training multiple models and combining their predictions. This paper examines combining an enhancement of the random subspace model of feature selection using fuzzy set similarity measures with different measures of evaluating feature subsets in the construction of an ensemble classifier. Experimental results show that in most cases a fuzzy set similarity measure paired with a feature subset evaluator outperforms the corresponding fuzzy similarity measure by itself and the learning process only needs to occur on typically about half the number of base classifiers since the features subset evaluator eliminates those feature subsets of low quality from use in the ensemble. In general, the fuzzy consistency index is the better performing feature subset evaluator, and inclusion maximum is the better performing fuzzy similarity measure.*

### *KEYWORDS*

*Feature selection, fuzzy set similarity measures, concordance correlation coefficient, feature subset evaluators, microarray data, ensemble learning.*

## 1. INTRODUCTION

Feature selection (FS) is an important task for classification in machine learning (ML) since it reduces dimensionality with respect to the feature dimension. Its objective is to find a possibly optimal feature subset of relevant features that reduces the data size and increases, or maintains, the overall performance measures such as accuracy and sensitivity on the results of the classification. Reducing the data size decreases data storage requirements and training times for learning algorithms and can improve visualization and interpretation of the learning results. There are three main approaches to feature selection: filter, wrapper and embedded methods. In this research, a filter method is used due to its advantages of typically being fast and not tuned for a given learner [1].

Feature selection methods have typically been performed by evaluating a candidate feature subset and searching through the feature space to find a better subset. Existing algorithms adopt various

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 73-86, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111407

measures to evaluate the quality of feature subsets [1][2][3]. The random subspace method (RSM) [4] for feature subset selection, however, does not use a search process. Instead, it randomly selects from an arbitrary sized subset of features that are ranked using an algorithm such as ReliefF [5], where ReliefF measures the relevance of a feature to the classification task. RSM techniques can be used to create an ensemble of base classifiers, each created from one of the randomly selected subsets of features [6]. Although such approaches are simple and fast, they do not consider possible correlations and dependencies that may exist between the features in the randomly selected subsets. In [7], ReliefF is used to rank the quality of the features and then the concordance correlation coefficient (CCC) [8] is used to group related features from the N top-ranked features into G disjoint subsets. An RSM-like approach is then used to randomly select a single feature from each of the G feature subsets instead of randomly selecting from all the top-ranked features. This process creates the feature subset for a single base classifier. This process is then repeated to create E base classifiers that are used in the ensemble. Extensions to that work examine the use of fuzzy set similarity measures (FSSM) along with the CCC to create the groups of related features and evaluate the difference in performance of the generated ensembles on four different datasets [9]. The FSSMs are modified to distance measure used in a hierarchical clustering process that creates the G feature subsets.

Some ML processes search through a space of feature subsets to find an optimal feature subset. They use feature subset evaluators to determine the quality of a feature subset. This paper further extends the research in [9] to employ different feature subset evaluators, not in a search process, but instead to assess the quality of each of the randomly generated feature subsets for use in base classifiers. Each feature subset evaluator is paired with a FSSM to determine its performance as compared to the corresponding FSSM alone. The hypothesis is that feature subset evaluators should reduce the number of base classifiers in the ensemble and improve the ensemble performance measures.

This research differs from the methods discussed in [10] which compares approaches used to create an ensemble of feature selectors and combine the produced feature subsets into one feature subset to be used in the machine learning process. In that research an ensemble of feature selectors is categorized as either homogenous or heterogenous. The homogenous selector uses the same feature selection method but on different training data subsets. The heterogenous feature selector ensemble uses a number of different feature selection methods but on the same training data.

Regardless of the type of feature selector ensemble, the resulting subsets of features must be aggregated into one feature subset. There are simple ways to approach this such as using the intersection or the union of the subsets of features [11]; however, these simple approaches may lead to a very restrictive set of features or to less reduction in the size of the set of features. A more sophisticated technique uses classification accuracy to combine the features produced by the various feature selectors [12]. This approach, however, is computationally expensive and may result in computational costs higher than that of the feature selection process.

The research presented here is similar to using homogeneous feature selectors in that the same modified RSM feature selector is used on different training datasets. It differs, however, since it does not combine the resulting subsets of features produced for each training dataset into one feature subset. Instead, each produced feature subset is used in the learning of a base classifier if it is of sufficient quality as determined by the feature subset evaluator. An ensemble is then created from those individual base classifiers learned using the feature subsets of sufficient quality. The ensemble is then applied on the test dataset, and the results of each of its base classifiers are combined using simple majority voting [13].

The paper is organized is as follows: Section II discusses the machine learning system which uses a combination of FSSMs with feature subset evaluators. Section III explains the FSSMs for grouping of similar features. The evaluators used to assess the quality of a feature subset are presented in section IV. This evaluation is over the feature set as a whole and differs from filtering methods applied to individual features. Section V discusses the experimental design and its parameters. Section VI presents the experimental results in terms of several views: 1) individual feature subset evaluators over FSSMs and datasets, 2) individual FSSMs over feature subset evaluators and datasets, 3) ensemble performance across datasets, 4) datasets across ensemble performance measures, and 5) highest ensemble performance values for pairs of feature subset evaluator and FSSM within datasets. Finally, section VII presents conclusions and possible future work.

## 2. Machine Learning Components

A machine learning system can have different structures and use a variety of methods. Here in this research the structure consists of 1) pre-process filtering, 2) a feature subset selection algorithm, 3) ensemble building algorithm, and 4) a learning algorithm. Figure 1 illustrates the system using four processes, which are described in the following four subsections.

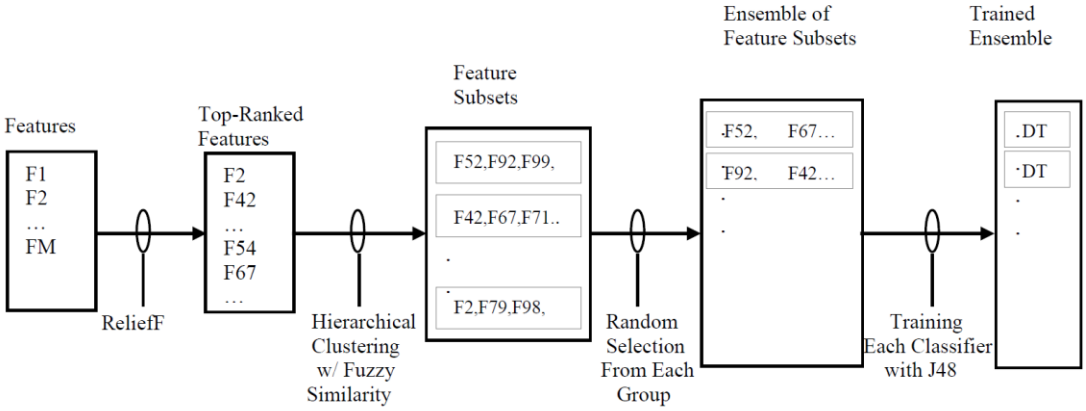


Figure 1. Machine Learning Process without Evaluators

### 2.1. Pre-processing Filtering

Filtering methods [14], also referred to as feature ranking methods, examine intrinsic properties of datasets to rank the features on their relevance to the classification task. This ranking is independent of the choice of learning algorithms.

ReliefF [5] is a well-known and often-used filtering method and is used to rank the features based on their numerical value. From this ranking, a specified number of top ranked features are selected as input to a feature subset selection algorithm; the others are discarded. Although ReliefF does provide a good way to assess the merit of individual features, it does not assess the merit of a *collection* of features. That is, a good feature set will include high-quality features in addition to having a diverse set of features; ReliefF is not designed to address this issue. The next step is to form subsets of these top-ranked features to use in training individual base classifiers.

### 2.2. Feature Subset Selection

Feature subset selection using feature subset evaluation produces candidate feature subsets based on a given strategy and can address feature redundancy in addition to feature relevance. A search strategy typically is used to search through feature subsets. Searching is time consuming due to feature subset generation and the evaluation of the feature subset. Methods to evaluate feature subsets are a distinguishing factor among feature selection algorithms using searching.

As done in [9], the top-ranked features undergo a hierarchical clustering algorithm to assign each feature uniquely into one of the G groups. When clustering, *distance* is defined using the specific FSSM used. During the clustering process, the two clusters that are merged are those that have smallest distance between the most distant members of the two individual clusters. When complete, the features that reside in a particular group can be viewed as being like the others in that group as defined by the FSSM, while being relatively dissimilar to the members of the other groups. The degree in which this is true is highly dependent on the underlying data and the value of G used.

This work does not use searching in the feature subset selection process. The RSM method, in combination with the grouping of related features, produces a feature subset by randomly selecting one feature from each of the G groups. In [9], no quality assessment is performed on each generated feature subset; each one is simply used in a base classifier to be trained for use in an ensemble. This current research instead uses different evaluators to assess the quality of a feature subset. If the feature subset's evaluation score is not sufficient, the feature subset is eliminated from use in a base classifier.

### 2.3. Ensemble Building

An ensemble can be built using a data partition, a feature partition, or hybrid approach [15]. Here, feature partitioning is used. Each feature subset of sufficient quality is associated with a base classifier. Each base classifier must be trained before used in the ensemble. An ensemble with multiple high-quality-only trained base classifiers is expected to have higher performance results and reduce learning times due to the elimination of low-quality feature subsets. The ensemble aggregates the predictions from its set of base classifiers using simple majority voting [13].

### 2.4. Machine Learning

Various machine learning algorithm with the training data can be used on a base classifier and its feature subset. Weka's J48 decision tree (DT) classifier [16][10] with default parameter settings is used for training the base classifiers. J48 is used for consistency with its use in [9].

## 3. Fuzzy Set Similarity Measures

Each feature is represented as a fuzzy set over the instances in the sample data sets. The feature values must be normalized to specify a degree of membership in [0, 1]. Similarity between the fuzzy sets representing each feature is determined using a FSSM. The fuzzy similarity measures are using during the hierarchical clustering process. In [9] the FSSMs used are presented in more detail. For completeness, they are briefly described here.

The *concordance correlation coefficient* (CCC) measures a bivariate relationship in terms of agreement between two values [8]. It differs from the Pearson correlation which measures the degree of linear relationship. CCC measures the degree to which pairs of values are close to the 45 degrees line of perfect concordance in a scatterplot. This line runs diagonally to the scatterplot. It is a very specific linear relationship, not just any linear relationship. A zero value indicates no agreement.

*Zadeh's consistency index*, also known as the sup-min or partial matching index [17], roughly estimates the similarity between two fuzzy sets by finding at what domain values they intersect and determines their similarity by taking the highest membership degree among their intersection points.

The *fuzzy Jaccard similarity measure* is a fuzzy extension of the Jaccard index [18] between two crisp sets. It replaces set cardinality with fuzzy set cardinality. It is the ratio between the fuzzy set cardinality of the intersection and that of the union.

A *fuzzy inclusion measure* determines how much one fuzzy set is included in another [17]. Another way to create a FSSM is to use a symmetric aggregation of the two directions of inclusion. The aggregation operators used are *average*, *minimum*, and *maximum*.

The *cosine* measure [19] views each fuzzy set as a vector in *n* dimensional space and computes the cosine of the angle between the two vectors. Because the feature values are values in the range [0, 1], the cosine can never be negative

## 4. FEATURE SUBSET SELECTION

Many feature selection methods contain two important aspects: evaluation of a candidate feature subset and searching through the feature space. The RSM approach does not use a search process but instead iteratively produces a randomly generated feature subset for a candidate base classifier to use in the ensemble. This current research incorporates the use of evaluation functions, referred to as *evaluators*, on the randomly generated feature subset. In [20], feature subset evaluators are classified into five categories: distance, information (or uncertainty), dependence, consistency**,** and classifier error rate. Evaluators in the classifier error rate category are referred to as wrapper methods [20]. Although wrapper methods produce high accuracy, due to their high computational cost, they are not considered here. The following describes the three evaluators used in this work.

*Interclass distance* (ICD) [2] in the distance category, also known as separability, divergence, or discrimination, is based on the assumption that instances of a different class should be distant in the instance space. Most often the distance measure *d* is in the Euclidean family:

$$\mathrm{ICD}(+,-) = \frac{1}{N_+N_-}\sum_{k_1}^{N_+}\sum_{k_2}^{N_-} d(x_{(+,k_1)}, x_{(-,k_2)}) \qquad (1)$$

where, + and - are the two class labels. $x_{(+,\ k1)}$ represents an instance $k_1$ of class +. $x_{(-,\ k2)}$ respresents an instance $k_2$ of class -. $N_+$ is the number of positive instances. $N_-$ is the number of negative instances. This formula is for two classes since the datasets in this study are binary classification problems. It takes the distance between each positive instance with each negative instance and sums over all possible pairs. Then the average over all the distances is taken.

*Maximal information compression index* (MiCi)[20] appears in both the information and dependence categories. An evaluator in the dependence category computes the dependence of a feature on other features. Its value measures the degree of redundancy of the feature. All evaluators in the dependence category can also be classified as information measures. MiCi measures the amount of error produced by reducing the pair of features (*x*, *y*) to a single feature. The greater the error means the less redundant are the two features. For features *x* and *y*, the formula is given as

$$MiCi(x,y) = \frac{1}{2}(var(x) + var(y) - \sqrt{(var(x) + var(y))^2} - 4(var(x) * var(y) * (1 - \rho(x,y)^2)) \tag{2}$$

*var(x)* is the variance of the values for feature *x* and similarly for *y*. *ρ(x,y)* is the covariance for the values of features *x* and *y*. This measure is performed for every pair of features in the feature subset, and the total is accumulated as a measure of error over the feature subset as a whole.

*Fuzzy consistency* (FC) is an adaptation of a crisp consistency measure [21] on a feature subset. Consistency for a feature subset F determines how many identical instances have the same class value for each group of identical instances, i.e., measures how consistent is the classification for each set of identical instances using the given F. Since this work only deals with binary classification problems, if there is a pattern A of identical instances, then for pattern A the crisp consistency is the maximum of either the positive class count or the negative class. The consistency formula for an identical pattern A is

$$C_F(A) = max_{k=+,-}\ [F_k(A)] \tag{3}$$

where $F_k$(A) is the number of instances in class k equal to A. The consistency rate is given as

$$CR(F) = \frac{\sum_{A \in S} C_F(A)}{|S|} \tag{4}$$

for each pattern A in the set of unique patterns S and |S| is the cardinality of the set S. Consistency measures rely on discrete-valued features where continuous features must first be discretized. To simplify measuring the consistency of the fuzzy instances, clustering is used to group instances into similar groups, i.e., |S| clusters, based on their feature membership values over features. Although the fuzzy instances in a group are not identical, the fuzzy consistency value is calculated for each similar group A as in Eq. 3. Fuzzy consistency rate is calculated as the average over all similar groups as in Eq. 4.

## 5. Experimental Design and Datasets

The research objective is to compare the effects of 1) the different FSSMs used to group features and 2) the evaluators used on the randomly generated features subsets from these groups on the performance when poorer feature subsets are eliminated. The results from the combinations of FSSMs and feature subset evaluators are compared to the results of just using FSSMs in creation of the base classifiers (i.e., without the evaluators). To perform this analysis, a systematic series of machine learning experiments were conducted, with the main control variables being the fuzzy similarity measure and the evaluator. The datasets used in these experiments are: breast, CNS, colon, and leukemia. The details of the data sets can be found in [9].

The objective in the previous research with FSSMs for grouping was to compare how the different similarity measures performed in creating ensemble classifiers. The reported results were based on finding the greatest performance values for accuracy, sensitivity, specificity, and F-measure for each fuzzy set similarity measure and dataset and the parameter values for which they occur. The input parameters for that research are S, N, G, and E. S is the FSSM. N is the number of top ranked features from ReliefF and ranged from 10 to 100 by an increment of 10. G is the number of clusters for grouping related features and ranged from 2 to 10 by an increment of 1. E is the number of base classifiers used to create the ensemble classifier and was fixed at 101. Note that the ensemble performance measures of accuracy, sensitivity, specificity, and F-measure may achieve their highest values at different N and G values for a FSSM and dataset.

In this current experiment, the objective is not to find the best achievable performance in the various ensemble performance measures. Instead, it is to compare performance results of a fuzzy set similarity used by itself in ensemble creation to those produced with the identical FSSM paired with each of the feature subset evaluators. This type of experiment can be used to assess the evaluators' effectiveness. To reduce the number of experiments and with this objective in mind, N is fixed at 50. This number was selected since across the experiments in the previous research, the majority of the highest performance measures were achieved at N ≤ 50. G is fixed at 10 since it was the highest value in the previous experiments and some highest performance values were achieved at 10. E is fixed at 101 where an odd E eliminates possible ties in majority voting.

The leave-one-out cross validation method is used in the experiments. An ensemble is to be created by independently learning with up to E base classifiers for each fold if the feature subset is of sufficient quality. Each feature subset is formed by randomly selecting one feature from each of the G feature subsets. The quality of the feature subset is then measured by using one of the three feature subset evaluators selected for the experiment. The acceptable quality level of a feature subset evaluator for it to be used in a base classifier is defined here for these experiments as

$$Quality(eval) = max_E(eval) - \frac{max_E(eval) - min_E(eval)}{2} \tag{5}$$

where *eval* is the evaluator used. The *max*$_E$*(eval)* is the maximum value of the subset evaluator *eval* over all the feature subsets for the potential base classifiers in the ensemble, and likewise, *min*$_E$*(eval)* is the minimum value. If a feature subset's quality value is greater than or equal to the acceptable quality level, that feature subset is included for training a base classifier in the ensemble. The maximum and minimum values for *eval* are over those values it produces for each generated feature subset. Only those feature subsets meeting the quality level are used to learn a base classifier. Each base classifier in the ensemble participates in a simple majority vote on the test sample.

## 6. EXPERIMENTAL PERFORMANCE RESULTS

Baseline experiments were conducted to obtain performance results from using only FSSMs for grouping the top-ranked ReliefF features. These results are then used for comparison with those results from the combination of FSSMs and feature subset evaluators used to eliminate poor feature subsets. Using FSSMs by themselves, E=101 base classifiers are always used. FSSMs combined with evaluators, the number of base classifiers may vary since the evaluation step can eliminate feature subsets that are determined to be of poor quality.

Tables I, II, III, and IV for the four datasets show comparisons on *accuracy, sensitivity, specificity* and *F-measure*, respectively. Each table shows each FSSM measure paired with a feature subset evaluator. The columns labeled "None" correspond to using the FSSM without any evaluator. These columns serve as the baseline. The other columns, ICD, MiCi, and FC, are the associated results when using the corresponding evaluator. For these columns, the values shown in bold for a value of FSSM and a feature subset evaluator pair meet or exceed those of the corresponding baseline result with no evaluator. Thus, bold entries are considered favorable and referred to as a "win." While matching the performance might be considered a "tie," it is viewed as a win because the same performance is obtained with fewer base classifiers.

The following analysis focuses on performance of the FSSMs and evaluators within a dataset based on its number of wins. The number of base classifiers used is discussed if there is no difference in wins analysis between two evaluators or between two FSSMs. Later analysis includes the actual number of base classifiers used when examining the highest ensemble performance measures.

Table I. Accuracy

| **FSSM** | **Breast** | | | | **CNS** | | | | **Colon** | | | | **Leukemia** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** |
| CCC | 0.633 | 0.653 | **0.714** | 0.673 | **0.517** | 0.483 | **0.500** | 0.500 | 0.806 | 0.790 | 0.823 | 0.839 | **0.986** | **0.972** | **0.972** | 0.972 |
| Cos | **0.673** | **0.673** | **0.673** | 0.673 | **0.450** | 0.433 | **0.450** | 0.450 | 0.677 | **0.694** | 0.677 | 0.694 | **0.875** | **0.875** | **0.875** | 0.875 |
| IncAve | **0.592** | **0.653** | **0.592** | 0.551 | **0.583** | **0.583** | **0.600** | 0.583 | 0.823 | 0.823 | 0.823 | 0.839 | 0.903 | 0.903 | **0.917** | 0.917 |
| IncMax | **0.612** | **0.571** | **0.592** | 0.510 | **0.617** | 0.567 | 0.583 | 0.600 | **0.855** | **0.871** | **0.871** | 0.855 | **0.972** | **0.972** | **0.986** | 0.972 |
| IncMin | **0.633** | **0.612** | **0.633** | 0.612 | 0.617 | **0.633** | **0.650** | 0.633 | **0.839** | 0.823 | **0.856** | 0.839 | **0.903** | **0.903** | **0.903** | 0.903 |
| Jaccard | **0.571** | **0.673** | **0.633** | 0.551 | **0.600** | **0.600** | **0.600** | 0.600 | **0.839** | **0.839** | 0.823 | 0.839 | 0.917 | **0.944** | 0.931 | 0.944 |
| Zadeh | 0.633 | 0.673 | 0.694 | 0.714 | **0.600** | 0.583 | 0.583 | 0.600 | **0.871** | **0.855** | **0.855** | 0.855 | 0.903 | 0.903 | **0.917** | 0.917 |
| **Wins** | 5 | 5 | 6 | n/a | 6 | 3 | 5 | n/a | 4 | 4 | 3 | n/a | 4 | 5 | 6 | n/a |

Table II. Sensitivity

| **FSSM** | **Breast** | | | | **CNS** | | | | **Colon** | | | | **Leukemia** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** |
| CCC | 0.600 | **0.680** | **0.680** | 0.640 | **0.095** | **0.048** | **0.048** | 0.048 | **0.900** | **0.900** | **0.900** | 0.900 | **1.000** | **0.979** | **0.979** | 0.979 |
| Cos | **0.680** | **0.680** | **0.680** | 0.680 | **0.238** | 0.143 | **0.190** | 0.190 | **0.775** | **0.775** | **0.775** | 0.775 | **0.897** | **0.894** | **0.894** | 0.894 |
| IncAve | **0.520** | **0.680** | **0.520** | 0.520 | **0.286** | 0.238 | **0.286** | 0.286 | **0.925** | **0.925** | **0.925** | 0.925 | 0.915 | 0.915 | **0.936** | 0.936 |
| IncMax | **0.640** | **0.560** | **0.560** | 0.480 | **0.238** | **0.238** | 0.095 | 0.143 | 0.900 | **0.925** | **0.925** | 0.925 | **0.979** | **0.979** | **1.000** | 0.979 |
| IncMin | **0.640** | 0.520 | **0.600** | 0.600 | **0.238** | 0.190 | **0.286** | 0.238 | **0.925** | **0.925** | **0.925** | 0.925 | 0.912 | **0.936** | 0.915 | 0.936 |
| Jaccard | **0.560** | **0.680** | **0.640** | 0.520 | 0.143 | 0.143 | 0.143 | 0.190 | **0.925** | **0.925** | 0.900 | 0.925 | 0.936 | 0.957 | 0.957 | 0.979 |
| Zadeh | 0.640 | 0.720 | 0.760 | 0.800 | **0.238** | 0.190 | **0.286** | 0.238 | **0.925** | 0.900 | **0.925** | 0.925 | 0.936 | 0.936 | 0.918 | 0.957 |
| **Wins** | 5 | 5 | 6 | n/a | 6 | 2 | 5 | n/a | 6 | 6 | 6 | n/a | 3 | 4 | 4 | n/a |

Table III. Specificity

| **FSSM** | **Breast** | | | | **CNS** | | | | **Colon** | | | | **Leukemia** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** |
| CCC | 0.667 | 0.625 | **0.750** | 0.708 | **0.744** | 0.718 | **0.744** | 0.744 | 0.636 | 0.591 | 0.682 | 0.727 | **0.960** | **0.960** | **0.960** | 0.960 |
| Cos | **0.667** | **0.667** | **0.667** | 0.667 | 0.564 | **0.590** | **0.590** | 0.590 | 0.500 | **0.545** | 0.500 | 0.545 | **0.840** | **0.840** | **0.840** | 0.840 |
| IncAve | **0.667** | **0.625** | **0.667** | 0.583 | **0.744** | **0.769** | **0.769** | 0.744 | 0.636 | 0.636 | 0.636 | 0.682 | **0.880** | **0.880** | **0.880** | 0.880 |
| IncMax | **0.583** | **0.583** | **0.625** | 0.542 | 0.821 | 0.744 | **0.846** | 0.846 | **0.773** | **0.773** | **0.773** | 0.727 | **0.960** | **0.960** | **0.960** | 0.960 |
| IncMin | **0.625** | **0.708** | **0.667** | 0.625 | 0.821 | **0.872** | **0.846** | 0.846 | **0.682** | 0.636 | **0.727** | 0.682 | **0.880** | **0.840** | **0.880** | 0.840 |
| Jaccard | **0.583** | **0.667** | **0.625** | 0.583 | **0.846** | **0.846** | **0.846** | 0.821 | **0.682** | **0.682** | **0.682** | 0.682 | **0.880** | **0.920** | **0.880** | 0.880 |
| Zadeh | **0.625** | **0.625** | **0.625** | 0.625 | **0.795** | **0.795** | 0.744 | 0.795 | **0.773** | **0.773** | **0.727** | 0.727 | **0.840** | **0.840** | **0.840** | 0.840 |
| **Wins** | 6 | 6 | 7 | n/a | 4 | 5 | 6 | n/a | 4 | 4 | 4 | n/a | 7 | 7 | 7 | n/a |

Table IV. F-Measure

| FSSM | Breast | | | | CNS | | | | Colon | | | | Leukemia | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** | **ICD** | **MiCi** | **FC** | **None** |
| CCC | 0.625 | **0.667** | **0.708** | 0.667 | **0.121** | 0.061 | **0.063** | 0.063 | 0.857 | 0.847 | 0.867 | 0.878 | **0.989** | **0.979** | **0.979** | 0.979 |
| Cos | **0.680** | **0.680** | **0.680** | 0.680 | **0.233** | 0.150 | **0.195** | 0.195 | 0.756 | **0.765** | 0.756 | 0.765 | **0.903** | **0.903** | **0.903** | 0.903 |
| IncAve | **0.565** | **0.667** | **0.565** | 0.542 | **0.324** | 0.286 | **0.333** | 0.324 | 0.871 | 0.871 | 0.871 | 0.881 | 0.925 | 0.925 | **0.936** | 0.936 |
| IncMax | **0.627** | **0.571** | **0.583** | 0.500 | **0.303** | **0.278** | 0.138 | 0.200 | 0.889 | **0.902** | **0.902** | 0.892 | **0.979** | **0.979** | **0.989** | 0.979 |
| IncMin | **0.640** | 0.578 | **0.625** | 0.612 | 0.303 | 0.267 | **0.367** | 0.313 | **0.881** | 0.871 | **0.892** | 0.881 | 0.925 | **0.926** | 0.925 | 0.926 |
| Jaccard | **0.571** | **0.680** | **0.640** | 0.542 | 0.200 | 0.200 | 0.200 | 0.250 | **0.881** | **0.881** | 0.867 | 0.881 | 0.936 | **0.957** | 0.947 | 0.958 |
| Zadeh | 0.640 | 0.692 | 0.717 | 0.741 | **0.294** | 0.242 | **0.324** | 0.294 | **0.902** | 0.889 | **0.892** | 0.892 | 0.926 | 0.926 | **0.938** | 0.938 |
| **Wins** | 5 | 5 | 6 | n/a | 5 | 1 | 5 | n/a | 3 | 3 | 3 | n/a | 3 | 5 | 5 | n/a |

## 6.1. Subset Evaluator Performance

First, the performance of evaluators is analyzed using the overall number of wins (*per column*) over all seven FSSMs and is discussed in terms of ensemble performance measures for each dataset. The following analysis includes numerous comparisons that, overall, show the evaluators can improve performance and reduce the size of the ensemble.

For the *breast* dataset, the *fuzzy consistency* evaluator is consistently the best or tied for the best in terms of the number wins against the baseline. For example, in *accuracy FC* has 6 wins. This is seen in the last row in Table 1 for the breast dataset. *ICD* and *MiCi* each have 5 wins. *FC* also performs better for the other performance measures with respect to the breast dataset.

For *CNS,* the *ICD* evaluator is best, or tied for the best, for *accuracy, sensitivity, and F-measure* as seen in the last row for the *CNS* column for Tables I, II, IV. *FC* is a close second for *accuracy. ICD* and *FC* are best for the *F-measure* as seen in last row for the *CNS* column of Table IV. *ICD,* however, is the worst for *specificity* as seen in Table III where *FC* is the best.

For the *colon* dataset, less variety exists between the evaluators. For example, in *accuracy*, both *ICD* and *MiCi* have 4 wins and *FC* only has 3 as seen in the last row in the *colon* column in Table I. For *sensitivity,* all evaluators have 6 wins, but *MiCi* might be judged the best if the least number of base classifiers required is considered. For *specificity* all three evaluators have 4 wins. For *F-measure,* all three evaluators have 3 wins.

For the *leukemia* data set, the *FC* evaluator is consistently the best, or tied for the best, with respect to all of the ensemble performance measures. *FC* has: 6 wins for accuracy in the leukemia column in Table I, 4 wins for sensitivity in Table II, 7 wins for specificity in Table III, and 5 wins for *F-measure* in Table IV. *MiCi* is nearly equal to *FC* but has only 5 wins for *accuracy.*

To summarize, *FC* is the best performer with respect to all datasets and ensemble performance measures. Its exceptions are for *sensitivity* in *CNS* and for *accuracy* in the *colon* dataset.

## 6.2. Fuzzy Set Similarity Performance

Next, the performance of FSSMs is analyzed using the overall number of wins (*per row)* for each FSSM across all the three evaluators and the four ensemble performance measures; therefore, there are a maximum of 12 wins for each data set. This analysis is done across all ensemble performance measures for a dataset since doing it per dataset for each ensemble performance measures provides only 3 cases to examine.

For the *breast* dataset, *Cos, IncAve, IncMax,* and *Jaccard* perform the best across all evaluators and all ensemble performance measures with 12 wins. For example, each row for *cos* is bold across all evaluators in each of the four tables for the *breast* dataset.

For the *CNS*, *IncAve* performs the best across all evaluators with 10 wins followed by *CCC* with 9 wins. For the *colon* dataset, *IncMax* and Zadeh perform the best across all evaluators with 10 wins.

For the *leukemia* dataset, *CCC, Cos,* and *IncMax* are the best performing FSSMs over all of the ensemble performance measures and evaluators with 12 wins.

## 6.3. Ensemble Performance Measures Across Datasets

Analysis across the ensemble performance measures can be examined across the 7 different FSSMs, each with 3 evaluators for a total of 21 cases. These 21 cases exist for each ensemble performance measures for each dataset. The range for the percentage of wins for each ensemble performance measure with respect to each dataset is presented.

For example, for *accuracy,* using pairs of FSSMs and evaluators, there are 16 wins for the *breast,* 15 wins for *leukemia* datasets, 14 wins for CNS, and 11 wins for colon. 16 wins corresponds to a 76% win rate (16/21). However, for accuracy in the colon dataset, there are only 11 wins or 52% of the 21 pairings. Thus, the range is from 52% (*colon*) to 76% ((*breast*). This result indicates that for most pairings of an evaluator with a FSSM the *accuracy* increases for all the datasets.

For *sensitivity,* the percentage of wins ranges from 52% (*Leukemia*) to 86% (*colon*). For *specificity,* the percentage of wins ranges from 57% (*colon*) to 100% (*leukemia*). For F-measure, the percentage of wins ranges from 43% (colon) to 76% (breast). Only for the *colon* dataset, is the percentage less than 50%. To summarize, pairing an evaluator with a FSSM has the most effect on *specificity* with the highest bottom and top values for the range. F-measure and accuracy have the lowest top range value at 76% and sensitivity has the lowest bottom range values at 43%.

## 6.4. Dataset Performance Across Ensemble Performance Measures

The range for the percentage of wins for each dataset with respect to each ensemble performance measure is presented. For the *breast* dataset, the percentage of wins ranges from 76% (*accuracy, sensitivity, F-measure*) to 90% (*specificity*). For *CNS,* the range is 52% (*F-measure*) to 71% (*specificity*). For the *colon* dataset, the range is 43% (*F-measure*) to 86% *(sensitivity*). For l*eukemia,* the range is 52% (*sensitivity*) to 100% (*specificity*). To summarize, the pairing of an evaluator with a FSSM has the most effect on *leukemia for* the highest top range; however, it does not have the highest bottom range. Breast has the highest bottom range and the second highest top range. Breast also has the smallest range where leukemia has the largest range. The smallest effect is on *CNS* since it has the lowest top range and almost the lowest bottom range but is second to the *colon* dataset.

## 6.5. Fuzzy Similarity and Evaluator Pairs with Highest Performance Measures

Finally, the performance of pairs of FSSMs and evaluators with respect to the highest values for each dataset and each ensemble performance measure is presented in Tables V, VI, VII, and VIII showing results for *accuracy, sensitivity, specificity,* and *F-measure,* respectively. The average number of base classifiers was not reported in previous tables. The number of base classifiers has been recorded for the experiments, but due to space limitations is only presented in Table V in the column labeled #BCs for the highest performing combinations.

Table V. Configurations With Best Accuracy

| Data | Acc. | FSSM | Eval | #BCs |
|---|---|---|---|---|
| Breast | 0.714 | CCC | FC | 59.6 |
| | | Zadeh | None | 101 |
| CNS | 0.650 | Jaccard | FC | 52.8 |
| Colon | 0.871 | IncMax | FC | 71.8 |
| | | | MiCi | 43.5 |
| | | Zadeh | ICD | 49.6 |
| Leukemia | 0.986 | CCC | ICD | 46.2 |
| | | IncMax | FC | 78.6 |

Table VI. Configurations With Best Sensitivity

| Data | Sens | FSSM | Eval | #BCs |
|---|---|---|---|---|
| Breast | 0.800 | Zadeh | None | 101 |
| CNS | 0.286 | IncAve | FC | 48.8 |
| | | | ICD | 50.5 |
| | | IncMin | FC | 52.8 |
| | | Zadeh | FC | 52.1 |
| Colon | 0.925 | IncAve | MiCi | 43.1 |
| | | | ICD | 54.3 |
| | | | FC | 75.3 |
| | | IncMax | MiCi | 43.5 |
| | | | FC | 71.8 |
| | | IncMin | MiCi | 38.7 |
| | | | ICD | 53.2 |
| | | | FC | 75.8 |
| | | Jaccard | MiCi | 42.4 |
| | | | ICD | 53.8 |
| | | Zadeh | ICD | 49.6 |
| | | | FC | 66.5 |
| Leukemia | 1.00 | CCC | ICD | 46.2 |
| | | IncMax | FC | 78.6 |

Table VII. Configurations With Best Specificity

| Data | Spec | FSSM | Eval | #BCs |
|---|---|---|---|---|
| Breast | 0.750 | CCC | FC | 59.6 |
| CNS | 0.872 | IncMin | MiCi | 30.1 |
| Colon | 0.773 | IncMax | FC | 71.8 |
| | | | MiCi | 43.5 |
| | | | ICD | 57.5 |
| | | Zadeh | MiCi | 38.9 |
| Leukemia | 0.960 | IncMax | FC | 78.6 |
| | | | MiCi | 34.0 |
| | | | ICD | 46.6 |
| | | | None | 101 |
| | | CCC | FC | 77.4 |
| | | | MiCi | 32.8 |
| | | | ICD | 46.2 |
| | | | None | 101 |

Table VIII Configurations With Best F-Measures

| Data | F-Meas | FSSM | Eval | #BCs |
|---|---|---|---|---|
| Breast | 0.741 | Zadeh | none | 101 |
| CNS | 0.367 | IncMin | FC | 52.8 |
| Colon | 0.902 | IncMax | FC | 71.8 |
| | | | MiCi | 43.5 |
| | | Zadeh | ICD | 49.6 |
| Leukemia | 0.989 | IncMax | FC | 78.6 |
| | | CCC | ICD | 46.2 |

With respect to evaluators, over all the datasets, *FC* has the highest number of instances for *accuracy* where it had the highest at 4. This can be seen by counting the number of rows in Table V that list FC as a top-ranked evaluator. For *sensitivity* (Table VI), FC has 8 followed by *ICD* with 6. For the *F-measure* (Table VIII), FC has at 3.

For *specificity* (Table VII), *MiCi* has the highest number of instances at 5 followed by *FC* at 4. When *MiCi* does produce one of the highest performance values, it always uses the least number of base classifiers.

Overall *FC* occurs at a highest performance value 19 times across all datasets: 4 times for *accuracy* and *specificity,* 8 times for *sensitivity*, and 3 times for *F-measure. FC* occurs with at least one FSSM for all ensemble performance measures over all datasets except for *F-measure* for *breast* and *specificity* for *CNS.*

With respect to FSSMs producing highest ensemble performance values, *IncMax* has its highest values occurring 8 times, 2 times for each ensemble performance measures and these were only for the *colon* and *leukemia* datasets *Zadeh* also has the highest values over all the ensemble performance measures with 8 occurrences, but only the *colon* dataset has all of the four ensemble performance measures at their highest values. All the other FSSMs occur 6 or fewer times with a highest performance value. Only *Cos* never produces a high for any ensemble performance measure.

Without evaluators (None), the FSSMs *CCC*, *IncMax* and *Zadeh* produce the highest, or tied for the highest, performance values. *Zadeh* produces the highest performance values for *sensitivity* and *F-measure* for the *breast* dataset and matches *CCC* for *accuracy* for the *breast* data set. *CCC* and *IncMax* without evaluators produce a highest *specificity* value for the *leukemia* dataset. When no evaluator is used with a FSSM, all 101 base classifiers are used, as shown in the columns labeled #BCS.

To summarize, for feature subset evaluators, overall *FC* produces the highest ensemble performance values. For FSSMs, *IncMax* and *Zadeh* produce the highest ensemble performance values. In terms of the number of wins as analyzed in Section 6.1 for ensemble performance measures, generally *FC,* regardless of the FSSM it is paired with, is the better performing feature subset evaluator. As analyzed in Section 6.2 for ensemble performance measures, *IncMax*, regardless of the feature subset evaluator it is paired with, is the better performing FSSM.

## 7. CONCLUSIONS

The research presented in this paper extends that in [9] where fuzzy set similarity measures (FSSMs) are used for grouping related features for an ML process. This current research employs the use of three feature subset evaluators in combination with seven FSSMs to examine their effects on the ensemble performance measures *accuracy, sensitivity, specificity* and *F-measure.*

First FSSMs are used to create groups of related features from the best ReliefF-ranked features. Next features are randomly selected from each group to produce a feature subset as in [9]. This random selection process occurs a fixed number of times to generate feature subsets to be associated with a fixed number of base classifiers for the ensemble. Typically, all base classifiers would be used in the ensemble. Instead, the quality of a feature subset associated with a base classifier is assessed using an evaluator. Those that have low quality are eliminated because they are likely to reduce the ensemble's performance.

Much research exists that discusses the use of feature subset evaluators in the search process of finding an optimal set of features for machine learning. Here three feature subset evaluators: interclass distance (*ICD*), maximal information compression index (*MiCi*), and fuzzy consistency (*FC*) are used, not in a search process, but to determine the quality of the feature subsets produced by the random subspace method of feature selection as applied to the feature groups formed using FSSMs. *FC* is an adaption of the crisp consistency measure which requires the discretization of feature values.

The experimental results showed that in most cases the FSSM paired with a feature subset evaluator outperforms the corresponding FSSM by itself, although it is acknowledged that it is difficult to know which combination will yield the most improvement. An added benefit is that the learning process only needs to occur on typically about half the number of base classifiers since the evaluator produces a quality assessment and those of low quality are eliminated from the ensemble.

From this study, in general the *FC* measure is the best performing feature subset evaluator paired with the FSSMs. As for FSSMs, in general *IncMax* paired with feature subset evaluators is the best performing for the colon and leukemia datasets. *CCC* and *Zadeh* with FC perform the best the breast and CNS datasets.

Future work will investigate other feature subset evaluators and the application of this pairing with FSSMs on other datasets. Initial experimental results also suggest that an aggregation of evaluators on a feature subset might present even higher quality feature subsets for an ensemble's base classifiers. The idea is that these higher quality feature subsets could further improve ensemble performance measures and reduce the number of needed base classifiers. In addition, a study to investigate possible relationships both between evaluators and ensemble performance measures and between evaluators and datasets might provide better insight to their use.

## REFERENCES

[1] Molina, L. C., Belanche, L., Nebot, A. (2002) "Feature Selection Algorithms: A Survey and Experimental Evaluation," *IEEE International Conference on Data Mining,* Maebashi City, Japan, Dec. 9 – 12.

[2] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2015) *Feature Selection for High-Dimensional Data,* Springer International Publishing AG, Switzerland.

[3] Wan, Cen (2019) *Hierarchical Feature Selection for Knowledge Discovery Application of Data Mining to the Biology of Ageing,* Springer Nature Switzerland AG.

[4] Ho, T. K. (1998) "The random subspace method for constructing decision forests," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 20 no. 8, pp. 832–844.

[5] Robnik-Sikonja, M. & Kononenko, I. (1997) "An adaptation of relief for attribute estimation in regression," in: D. H. Fisher 635 (Ed.), *Fourteenth International Conference on Machine Learning,* Morgan Kaufmann, pp. 296–304.

[6] Kuncheva, L. I., Rodríguez, J. J., Plumpton, C. O., Linden, D. E. J., & Johnston, S. J. (2010) "Random Subspace Ensembles for fMRI Classification," *IEEE Transactions on Medical Imaging,* Vol. 29, No. 2, pp. 531- 542.

[7] Chaudhury, B., Goldgof, D. B., Hall, L. O., Gatenby, R. A., Gillies, R. J., & Drukteinis, J. S. (2015) “Correlation based random subspace ensembles for predicting number of axillary lymph node metastases in breast dce-mri tumors,” *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2164–2169.

[8] Bland, J. M. & Altman, D. (1986) “Statistical methods for assessing agreement between two methods of clinical measurement,” *The Lancet*, vol. 327 no. 8476, pp. 307–310.

[9] Cross, V., Zmuda, M., Paul, R., & Hall, L. O. (2020) “Fuzzy Set Similarity for Feature Selection in Classification”, *2020 International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 19 – 24, Glasglow, United Kingdom.

[10] Bolón-Canedo, Verónica & Alonso-Betanzos, Amparo (2019) “Ensemble for feature selection: Review and Trends,” *Information Fusion*, Vol 52, pp. 1-12.

[11] Álvarez-Estévez, Diego, Sánchez-Maroño, Noelia, Alonso-Betanzos, Amparo, & Moret-Bonillo, Vicente (2011) “Reducing dimesionality in a database EEG sleep arousals,” *Expert Systems with Applications*, 38(6), pp. 7746-7754.

[12] Morán-Fernández, L., Bolón-Canedo, V., & Alonso-Betanzos, A. (2017) “Centralized vs. distributed feature selection methods based on data complexity measures,” *Knowl. Based Syst.* 117 pp. 27–45.

[13] Liu, H., Liu, L., & Zhang, H. (2010) “Ensemble gene selection by grouping for microarray data classification,” *Journal of Biomedical informatics*, 43 (1) pp, 81–87.

[14] Duch, W. (2006) “Filter Methods,” in *Feature Extraction Foundations and Applications*, Eds. I. Guyon, M. Nikravesh, S. Gunn, L. Zadeh, Berlin: Springer Berlin Heidelberg.

[15] L. Rokach (2010), “Ensemble-based classifiers,” *Artif Intell Rev*, vol. 33, pp. 1–39.

[16] Holmes, G., Donkin, A., & Witten, I. H. (1994) “Weka: A machine learning workbench,” *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994 Web link: https://www.cs.waikato.ac.nz/ml/weka/

[17] Dubois, D. & Prade, H. (1982) “A unifying view of comparison indices in a fuzzy set-theoretic framework,” in R Yager Ed. *Fuzzy Set and Possibility Theory: Recent Developments*, Pergamon Press, New York, NY pp. 3-13.

[18] P. Jaccard (1912) "The distribution of the flora in the alpine zone", *New Phytologist*, vol. 11, pp. 37–50.

[19] Han, Jiawei, Kamber, M., & Pei, Jian (2012) *Data Mining: Concepts and Techniques*, 3rd Ed. Morgan Kaufmann, Burlington, MA.

[20] Mitra, P., Murthy, C.A., & Pal, S.K. (2002) “Unsupervised Feature Selection Using Feature Similairty,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 3.

[21] Dash, M. & Liu, Huan (2003) “Consistency-based search in feature selection,” *Artificial Intelligence*, Vol. 151, pp. 155-176.

## AUTHORS

**Dr. Valerie Cross** is an Associate Professor in the Computer Science and Software Engineering department at Miami University in Oxford, OH. She earned a B.S in Computer Science and a B.S. in Statistics from West Virginia University, a Masters in Computer Science at the University of Colorado, Boulder and a PhD in Computer Science from Wright State University. Her research interests include fuzzy set theory and approximate reasoning, ontology alignment, ontologies in biomedical and bioinformatics applications and the use of fuzzy set theory in machine learning.

**Dr. Zmuda** is an Associate Professor in the Computer Science and Software Engineering department at Miami University in Oxford, OH. He earned a B.S in Computer Science and a B.S. in Mathematics from Eastern Michigan University and a M.S. and Ph.D. in Computer Science and Engineering at Wright State University. His research interests include the application of AI techniques such as fuzzy set theory and optimization to problems in medicine and virtual reality.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# A DEEP LEARNING BASED APPROACH TO ARGUMENT RECOMMENDATION

Guangjie Li, Yi Tang, Biyi Yi, Xiang Zhang and Yan He

National Innovation Institute of Defense Technology, Beijing, China

## *ABSTRACT*

*Code completion is one of the most useful features provided by advanced IDEs and is widely used by software developers. However, as a kind of code completion, recommending arguments for method calls is less used. Most of existing argument recommendation approaches provide a long list of syntactically correct candidate arguments, which is difficult for software engineers to select the correct arguments from the long list. To this end, we propose a deep learning based approach to recommending arguments instantly when programmers type in method names they intend to invoke. First, we extract context information from a large corpus of open-source applications. Second, we preprocess the extracted dataset, which involves natural language processing and data embedding. Third, we feed the preprocessed dataset to a specially designed convolutional neural network to rank and recommend actual arguments. With the resulting CNN model trained with sample applications, we can sort the candidate arguments in a reasonable order and recommend the first one as the correct argument. We evaluate the proposed approach on 100 open-source Java applications. Results suggest that the proposed approach outperforms the state-of-the-art approaches in recommending arguments.*

## *KEYWORDS*

*Argument recommendation, Code Completion, CNN, Deep Learning*

## 1. INTRODUCTION

Code completion is one of the most widely used Eclipse features by developers. It may automatically completes the rest part of an expression or statement when developers type in the first several characters, which helps speed up coding and as a result the whole process of software development.

Argument recommendation is a special kind of code completion. Most of the mainstream IDEs recommend actual arguments for method calls when developers type in method names. However, such IDEs only provide a long list of candidate arguments according to the corresponding types of formal parameters, which may take a long time for developers to select the correct one from the list of type compatible candidate arguments.

To facilitate the process of development, a few approaches have been proposed to recommend arguments. Zhang et al. [4] recommend arguments for method invocations based on the nearest k usages of them. Raychev et al. [6] recommend arguments for method invocations based on statistical language model. Such approaches can only work well for methods with richful

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 87-93, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111408

invocation histories, however, a large number of methods in practice have less or no invocation history before the current method call, consequently the state-of-art approaches cannot be used to recommend arguments for such method call. For example, according to Li et al. [1], nearly one half of method invocations are non-API method invocations, i.e., methods defined within the projects.

To this end, in this paper we propose a deep learning based approach to recommend arguments for both API method invocations and non-API method invocations based on features extracted from the context of method invocations. First, we extract context information for each method invocation from a large number of open-source applications, which involves method names, formal parameters, actual arguments, type compatible variable names, type compatible and visible method names. Second, we perform natural logarithm transformation and normalization to the dataset. Third, we feed the preprocessed dataset to a specially designed Convoluntional Neural Network (CNN) so as to learn the general rules of mapping candidate arguments into parameters. Fourth, we rank the candidate arguments according to the predicted probabilities of being actual arguments in descending order, and recommend the first one as the actual argument. Evaluations on 100 open-source applications suggest that the proposed approach outperforms the state-of-art approaches in recommending arguments for method invocations.

This paper makes the following contributions:

- To the best of our knowledge, it is the first one in recommending and ranking arguments for methods.

- Evaluations on real-world open-source applications suggest that the proposed approach outperforms the state-of-the-art approach in recommending arguments for method invocations.

- We exploit natural language processing techniques to mine lexical similarities embedded in software identifiers, and exploit word embedding and deep learning techniques to mine semantic similarities between related program entities.

The rest of the paper is organized as follows. Section 2 describes the proposed approach. Section 3 presents an evaluation of the proposed approach on open-source applications. Section 4 presents related works. Section 5 provides conclusions.

## 2. APPROACH

### 2.1. Overview

The rationale of the approach is that we can select correct argument from a list of syntactically candidate arguments based on method invocation contexts. Consequently, we train an CNN (convoluntional neural network) model with training data, i.e., actual arguments and their context from open source applications, and then rank and recommend correct arguments for new call sites based on the resulting neural network. An overview of the proposed approach is presented in Figure 1.

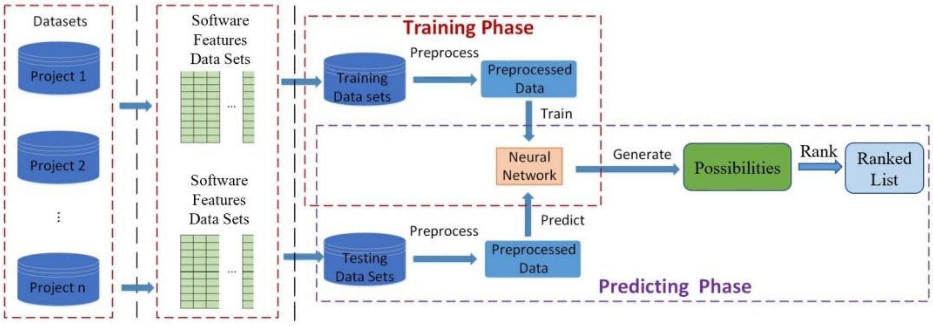


Fig. 1. Overview of the proposed approach.

For a given argument in a method invocation, we exploit the JDT (Eclipse plug-in tool) to parse the ASTs of Java files statistically and extract the following context information:

- Ar: the actual argument name.
- Fn: the method name.
- Par: the formal parameter name.
- Lvs: all variables visible and type compatible.
- Mvs: all methods visible and type compatible.

where Lvs and Mvs are a list of candidate arguments, respectively. It should be noted that, for a given recommendation position, we only consider those expressions as candidate arguments when choosing them as the actual argument will not induce syntactical errors.

## 2.2. Data Preprocessing

To rank and recommend candidate arguments, we need to preprocess data extracted from program and feed them into the CNN, which involves the following steps. First, we split each identifier into a sequence of tokens by exploiting underscore and capital letter as separators. Second, we exploit Word2Vec [3] to embed the token sequences of each identifier into numerical vectors. Third, if the candidate arguments are more than five, we only remain the top five candidates who are lexical similar to the parameters based on computing Jaccard similarity [13].

## 2.3. CNN-Based Architecture

The architecture of the neural network for argument recommendation is presented in Figure 4. The model consists of five input layers, five convolutional layers, two fully connected layers, and one output layer. Preprocessed data are divided into groups and each group is input to the corresponding convolutional neural network respectively. We set the input dimensions, output dimensions, kernel initializer and activation function for each layer as follows:

- Convolutional layers: kernel_initializer =glorot_uniform, activation function = Softmax, pooling =MaxPooling1D, and dropout = 0.25.

- First fully connected layer: output_dim = 32, activation function =Softmax, dropout = 0.5.

- Second fully connected layer: out_dim =5.

The output of the CNN-based network is the possibilities of candidates as the actual argument for a given recommendation requirement, and the proposed approach selects the one with highest possibility as the recommended one. Each CNN layer is forwarded to a flatten layer, the merge layer merges the outputs of the flatten layers as a vector, and feed them into the fully connected dense layer. Finally, the dense layer outputs the prediction for each instance.

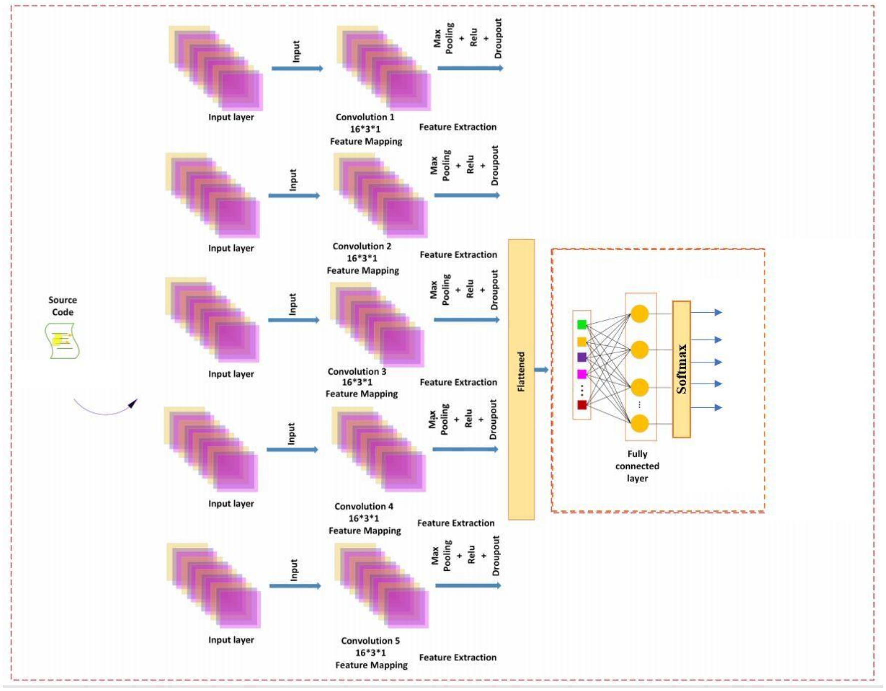

Source Code
Input layer
Input
Convolution 1 16*3*1 Feature Mapping
Max Pooling + Relu + Droupout
Feature Extraction
Input layer
Input
Convolution 2 16*3*1 Feature Mapping
Max Pooling + Relu + Droupout
Feature Extraction
Input layer
Input
Convolution 3 16*3*1 Feature Mapping
Max Pooling + Relu + Droupout
Feature Extraction
Input layer
Input
Convolution 4 16*3*1 Feature Mapping
Max Pooling + Relu + Droupout
Feature Extraction
Input layer
Input
Convolution 5 16*3*1 Feature Mapping
Max Pooling + Relu + Droupout
Feature Extraction
Flattened
Softmax
Fully connected layer

Figure 2. Overview of the CNN-based classifier

## 3. Evaluation

This section specifies the setup of the evaluation, research questions, and metrics employed to evaluate the performance of the proposed approach. To evaluate the state-of-art argument recommendation approach, we select the similarity-based approach for comparison because of the following reasons. First, the similarity-based approach is the most recently proposed argument recommendation approach for method invocations. Second, the similarity-based approach does not rely on richful invocation history of the invoked method, which is similar to our approach. Third, the source code of the similarity-based approach is publicly available, which makes it easy to repeat their experiment.

We evaluate the proposed approach on real-world applications. We search for most popular (stars) 100 open-source Java applications from GitHub as the subject applications, select 90 of the resulting applications as training dataset, and the left 10 applications as testing dataset.

### 3.1. Research Questions

The evaluation investigates the following research questions:

- RQ1: Does the proposed approach outperform the state-of-art approach in recommending arguments for method invocations?

- RQ2: How long does it take to train the neural network model, and how long does it take to generate recommendation for a given method invocation?

- Research question RQ1 validates the performance of the proposed approach. Answering this question may reveal whether deep learning techniques outperform fine-grained heuristic rules in mining semantic relationships.

- Research question RQ2 reveals the efficiency of the proposed approach. Answering this question may help to validate whether the proposed approach can be applied in practice.

## 3.2. Metrics

To measure the performance of the approaches, we define precision and recall as follows:

$$\text{precision} = \frac{\text{Num}_{accepted}}{Num_{recommended}} \tag{1}$$

$$\text{recall} = \frac{\text{Num}_{accepted}}{Num_{\text{tested}}} \tag{2}$$

$$\text{F1} = 2* \frac{\text{precision} * \text{recall}}{precision + recall} \tag{3}$$

where $\text{Num}_{\text{accepted}}$ is the number of correct recommendations, $\text{Num}_{\text{recommended}}$ is the number of generated recommendations, and the $\text{Num}_{\text{tested}}$ is the number of arguments extracted from the object applications.

## 3.3. Results

Evaluation results are presented in Fig 3. From this figure, we observe that the proposed approach significantly outperforms existing approaches in recommending arguments.

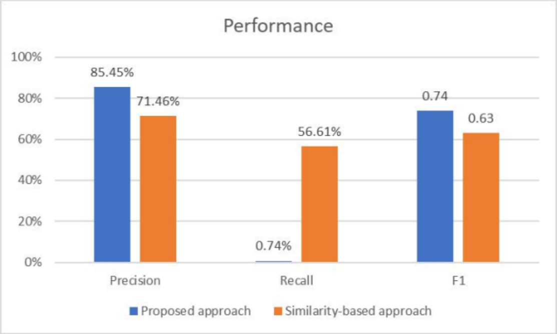


Figure 3. Evaluation Results

We also evaluate time efficiency of the proposed approach in training and testing phrase, respectively. Training task is conducted on a special workstation with the following configuration: 2.0GHz Intel Xeon E5-2683 processor, 64GB RAM, TITAN Xp GPU, with Linux installed. Testing is conducted on a personal computer with the following configuration: Intel Core i7-6700 CPU 3.4 GHz, 16 GB RAM, with Windows 7 installed. Evaluation results suggest that it takes around 89 minutes to train the CNN model and around 11 milliseconds on average to make recommendation for each argument requirement.

## 4. Related Works

N-gram language model, rooted in statistical natural language processing, has been proved to be successful in capturing the repetitive and predictable regularities of source code [2]. Consequently, a series of n-gram based approaches have been proposed to predict the naturalness of code. Hindle et al. [2] recommend the next code token based on the preceding n tokens by training n-gram learning model. Allamanis et al. [6] exploit n-gram models to recommend variable names, method names and class names. Tu et al. [4] exploit the localness of source code in recommending the next token by adding a cache component to the n-gram

model, i.e, assigning a higher probability to tokens occurred in the source file where the n-gram model based prediction is applied. Hellendoorn et al. [11] model and complete source code based on a nested and cached n-gram based approach. Based on the naturalness and localness of source code, they assign a higher probability to tokens most recently used by adding a cache mechanism to the n-gram model. Raychev et al. [7] exploit statistical language models and conditional random fields in predicting local variable names for JavaScript applications. Nguyen et al. [5][8] exploit graphic probability models to recommend the next API method call.

Neural network based approaches are also proposed to improve code completion. White et al. [9] exploit deep learning to model software and illustrate that deep learning based approach is more effective than n-gram based one. Murali et al. [10] train a deep learning based model to generate code fragment for program sketches that heavily dependent on APIs. Most recently, Liu et al.[12] propose a similarity-based approach to recommend arguments. They just select the candidate who has the greatest lexical similarity with the corresponding parameter as the recommended argument.

## 5. Conclusions

In this paper, we propose a deep learning based approach to rank candidate arguments and recommend actual argument for method invocations. By statistically parsing Java source files from open-source applications, we extract each actual argument and the corresponding context information from each method invocation, represent them in vectors, and feed them into a specially designed CNN classifier to learn the rules of selecting correct arguments. Evaluations on 100 open-source applications suggest that the proposed approach outperforms the state-of-art approaching in recommending arguments. The insight of the approach is that deep learning techniques can effectively learn the semantic similarity between related software entities, and they can be used to facilitate software engineering task.

## References

[1] Guangjie Li , H Liu, Ge Li , et al. LSTM-based argument recommendation for non-API methods[J]. Science China (Information Sciences), 2020(9).

[2] A. Hindle, E. T. Barr, Z. Su, M. Gabel, and P. Devanbu, “On the naturalness of software,” in 2012 34th International Conference on Software Engineering (ICSE), June 2012, pp. 837–847.

[3] T. Mikolov, K. Chen, G. Corrado, and J. ean, “Efficient estimation of word representations in vector space,” Computer Science, 2013.

[4] Tu Z, Su Z, Devanbu P. On the localness of software. In: Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering. New York: ACM, 2014. 269–280

[5] A. T. Nguyen and T. N. Nguyen, “Graph-based statistical language model for code,” in Proceedings of the 37th International Conference on Software Engineering - Volume 1, ser. ICSE’15. Piscataway, NJ, USA: IEEE Press, 2015, pp. 858–868. [Online]. Available: http://dl.acm.org/citation.cfm?id=2818754.2818858

[6] M. Allamanis and C. Sutton, “Mining source code repositories at mas sive scale using language modeling,” in 2013 10th Working Conference on Mining Software Repositories (MSR), May 2013, pp. 207–216.

[7] V. Raychev, M. Vechev, and E. Yahav, “Code completion with statistical language models,” in Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, ser. PLDI ’14. New York, NY, USA: ACM, 2014, pp. 419–428. [Online]. Available: http://doi.acm.org/10.1145/2594291.2594321

[8] T. T. Nguyen, H. V. Pham, P. M. Vu, and T. T. Nguyen, “Learning api usages from bytecode: A statistical approach,” in Proceedings of the 38th International Conference on Software Engineering, ser. ICSE’16. New York, NY, USA: ACM, 2016, pp. 416–427. [Online]. Available: http://doi.acm.org/10.1145/2884781.2884873

[9] White M, Vendome C, Linares-Vasquez M, et al. Toward deep learning software repositories. In: Proceedings of the 12th Working Conference on Mining Software Repositories. Piscataway: IEEE Press, 2015. 334–345

[10] Murali V, Qi L, Chaudhuri S, et al. Neural sketch learning for conditional program generation. 2017. ArXiv: 1703.05698

[11] Hellendoorn V J, Devanbu P. Are deep neural networks the best choice for modeling source code? In: Proceedings of Joint Meeting on Foundations of Software Engineering, 2017. 763– 773

[12] Liu H, Liu Q, Staicu C A, et al. Nomen est omen: exploring and exploiting similarities between argument and parameter names. In: Proceedings of the 38th International Conference on Software Engineering. New York: ACM, 2016. 1063–1073

[13] W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg, “A comparison of string distance metrics for name-matching tasks,” in Proc. Workshop Inf. Integr. Web (IIWeb), 2003, pp. 73–78.

## AUTHORS

**Guangjie Li** received the B.S. and M.S. degrees in educational technology from Shen Yang Normal University in 2002 and 2005, respectively. She received her Ph.D. degree in computer science from and technology from Beijing Institute of Technology in 2020. She is is interested in software quality and software evolution.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# DEALING CRISIS MANAGEMENT USING AI

Yew Kee Wong

School of Information Engineering, Huang Huai University, Henan, China.

## ABSTRACT

*Artificial intelligence has been a buzz word that is impacting every industry in the world. With the rise of such advanced technology, there will be always a question regarding its impact on our social life, environment and economy thus impacting all efforts exerted towards sustainable development. In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets for different industries and business operations. Numerous use cases have shown that AI can ensure an effective supply of information to citizens, users and customers in times of crisis. This paper aims to analyse some of the different methods and scenario which can be applied to AI and big data, as well as the opportunities provided by the application in various business operations and crisis management domains.*

## KEYWORDS

*Artificial Intelligence, Big Data, Business Operations, Crisis Management*

## 1. INTRODUCTION

Artificial intelligence (AI) is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think. AI is accomplished by studying how human brain thinks, and how people learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems [1]. AI is a science and innovation based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering. A major thrust of AI is in the development of computer functions associated with human intelligence, for example, reasoning, learning, and problem solving. Out of the following areas, one or multiple areas can contribute to build an intelligent system [2]. This paper aims to analyse some of the use of big data for the AI development and its applications in various business operations and crisis management.

## 2. WHAT IS BIG DATA

The Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. A buzzword that is used to describe immense volumes of data, unstructured, structured and semi-structured, big data can inundate a business on a day-to-day basis. Big data is used to analyse insights, which can lead to better decisions and strategic business moves [3]. The definition of big data: "Big data is high-volume, and high-velocity or high-variety information assets that demand

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 95-102, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111409

cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." The characteristics of Big Data are commonly referred to as the four Vs:

## Volume of Big Data

The volume of data refers to the size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Asia.

## Velocity of Big Data

Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Instagram messages or Wechat posts.

## Variety of Big Data

Variety makes Big Data really big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.

## Veracity of Big Data

Veracity refers to the quality of the data that is being analysed. High veracity data has many records that are valuable to analyse and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labelled Big Data [4].

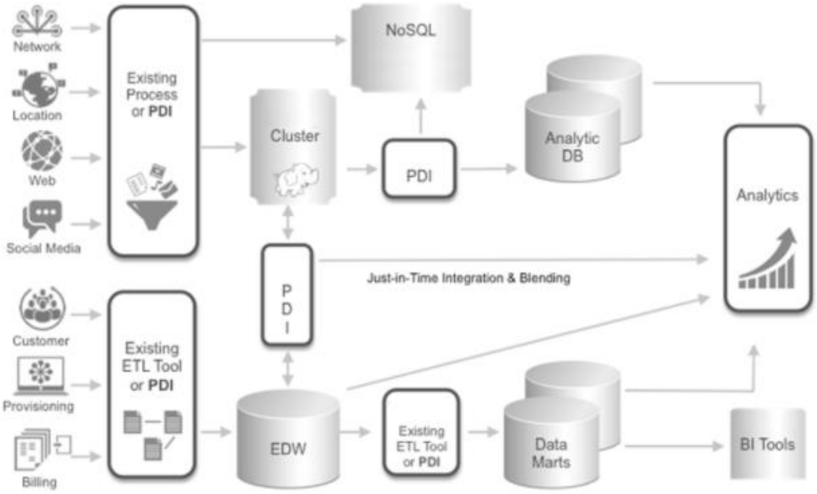


Figure 1. Big Data Architecture. (arccil.com)

## 2.1. Types of Big Data

There are 3 types of big data; unstructured data, structured data and semi-structured data.

**Unstructured Data:**

Any data with unknown form or the structure is classified as unstructured data.

**Structured Data:**

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

**Semi-Structured Data:**

Semi-structured data can contain both the forms of data.

Dealing with unstructured and structured data, data science is a field that comprises everything that is related to data cleansing, preparation, and analysis. Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data [5]. This umbrella term includes various techniques that are used when extracting insights and information from data.

**Big data benefits:**

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data, which means a completely different approach to tackling problems.

### 2.2. What is Big Data Analytics

Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models [6] [7]. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

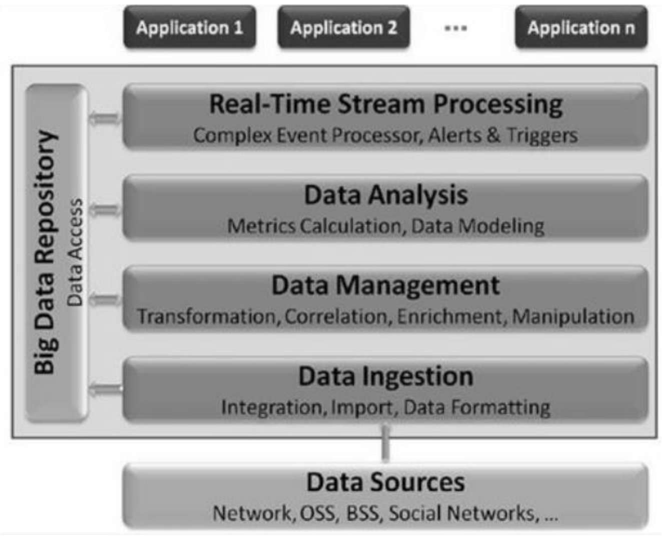

Application 1
Application 2
...
Application n
Big Data Repository
Data Access
Real-Time Stream Processing
Complex Event Processor, Alerts & Triggers
Data Analysis
Metrics Calculation, Data Modeling
Data Management
Transformation, Correlation, Enrichment, Manipulation
Data Ingestion
Integration, Import, Data Formatting
Data Sources
Network, OSS, BSS, Social Networks, ...

Figure 2. Big Data Analytics Architecture.

## 3. USING AI IN SENSITIVE BUSINESS OPERATIONS

The artificial intelligence rules define the way the online learning system assigned learning materials and exercises for the learner to follow [8]. These are the basic rules which we have carry out in our experiments, in which we find it effective in improving the learners understanding.

### 3.1. Financial Industry

Artificial intelligence (AI), along with other financial technology (fintech) innovations, are significantly changing the ways that financial business are being run, especially in the fields like trading and insurance, leading the traditional financial industry into a new era [9].

### Robots Replacing Humans

Back in 2000, Goldman Sach's New York headquarters employed 600 traders, buying and selling stock on the orders of the investment bank's clients. Today there are just two equity traders left, as automated trading programs have taken over the rest of the work. Meanwhile, BlackRock, the world's biggest money manager, also cut more than 40 jobs earlier this year, replacing some of its human portfolio managers with artificially intelligent, computerized stock- trading algorithms. Those two big companies are not the only financial institutions replacing human jobs with robots. By 2025, AI technologies will reduce employees in the capital markets by 230,000 people worldwide, according to a report by the financial services consultancy Opimas [10].

Big new frontiers are only just beginning to opening up in fintech from AI, block chain and robotics to biometrics, augmented reality and cybersecurity. Among all the fintech innovations, the prospect of the block chain has the highest expectation. The block chain will change the way people store information, which is real, spreading fast and cross-border, and its 'de-centric' feature will allow everyone to know what other people are doing. The application of block chain in finance will once again bring about a revolutionary impact on the industry, just like AI does.

## 3.2. Health Industry

The Artificial intelligence (AI) is reshaping operations across industries. Arguably, healthcare is where these changes are poised to make the biggest impact – optimizing uptime and availability of the treatment solutions. Using AI-powered tools capable of processing large amounts of data and making real-time recommendations, healthcare organizations are learning they can reduce administrative waste in a number of areas, from medical equipment maintenance to hospital bed assignments [11].

Artificial intelligence is reinventing and reinvigorating modern healthcare through technologies that can predict, comprehend, learn and act. The ability of AI to transform clinical care has received widespread attention, but the technology’s potential extends beyond patient care to processes across the spectrum of healthcare operations. In healthcare and other industries that depend on reliable equipment performance, few things are more disruptive than unexpected outages. These unplanned stops create costly emergency situations, such as extended downtime, rush delivery of parts and overtime to repair the equipment.

Facing pressure to improve profitability and efficiency, many healthcare organizations are turning to emerging technologies like AI and big data analytics to improve upon existing maintenance operations. Until recently, maintenance typically involved either reacting to an unexpected problem or adhering to a preventive maintenance schedule, which can sometimes result in unnecessary maintenance. line.

## 3.3. Manufacturing Industry

AI is core to manufacturing's real-time future. Real-time monitoring provides many benefits, including troubleshooting production bottlenecks, tracking scrap rates, meeting customer delivery dates, and more. It's an excellent source of contextually relevant data that can be used for training machine learning models. Supervised and unsupervised machine learning algorithms can interpret multiple production shifts' real-time data in seconds and discover previously unknown processes, products, and workflow patterns [12].

The manufacturing industry has exploited the use of AI technology, and in particular knowledge-based systems, throughout the manufacturing lifecycle. This has been motivated by the competitive challenge of improving quality while at the same time decreasing costs and reducing design and production time. Just-in-time manufacturing and simultaneous engineering have further required companies to focus on exploiting technology to improve manufacture planning and coordination, and on providing more intelligent processing in all aspects of manufacturing. The objective is to improve quality, to reduce costs, and to speed up the design and manufacturing process.

# 4. USING AI IN CRISIS MANAGEMENT

## 4.1. Extreme Weather Forecast

According to the UN Office for the Coordination of Human Affairs, in 2016 over 100 million people were affected by natural disasters including earthquakes, hurricanes and floods. Technology has a vital role to play in providing the appropriate situational awareness that then shapes practical, life-saving decisions for effective crisis management. These decisions may involve the evacuation of the most dangerous areas after an earthquake, or explore tactical options about how and where to position critical resources like medicine, food, clean water and shelter. Through utilising the data tweeted and texted by citizens in a crisis zone, rescuers have access to the knowledge needed to devise a strategy for immediate rescue attempts and for longer term help [13].

Issues can arise, however, due to the volume of available data, and high-quality filtering systems are needed to avoid using inaccurate data that could misdirect humanitarian aid, potentially wasting time, resources, and human trust in the system. Humanitarian responders may, understandably, question the specificity of information, therefore, building their trust and encouraging uptake of AI technology is a socially meaningful endeavour; without this, a system is unlikely to be adopted in the field. Machine learning, understood as the refinement of how AI ‘learns’ to use algorithms and other data, offers a solution to detecting key information taken from social media messages. Hence, researchers are focusing efforts on improving how the millions of messages are sifted by algorithms to overcome inaccuracy, ensuring that only the most important data is identified and shared.

## 4.2. Man-Made Environmental Disaster

The case of BP oil spill in 2010 provides an important example for understanding how these principles are valued by public opinion in a crisis situation, and how the communication actions by a corporation in this type of circumstances might have long-term effect on the brand image of the organization. On April 20, 2010, a BP’s Deepwater Horizon oil rig exploded, causing what has been called the worst environmental disaster in U.S. history and taking the lives of 11 rig workers. For 87 straight days, oil and methane gas spewed from an uncapped well-head, 1 mile below the surface of the ocean. The federal government estimated 4.2 million barrels of oil spilled into the Gulf of Mexico [14].

The accumulation of unsafe supervisory action had resulted in risk levels substantially increasing. Not only were risks increasing, but they were also incrementally becoming more aggressive in nature. For instance, one of the first acts of unsafe supervision is illustrated when BP neglected its responsibility of ensuring safety protocols were carried out after the completion of the Macondo Well. This was a major mistake on BP’s part, violating safety protocols which may have identified the issues present with the cementing of the well. Should these issues have been identified sooner, the likeliness of the crisis happening would potentially be slim. In addition to this, there was also very little supervision during and after works were carried out. This can be attributed to the aforementioned organisational restructuring which created much confusion regarding who was accountable for the assurance of safety [15].

## 4.3. Natural Disaster

Researchers have found that AI can be used to predict natural disasters. With enormous amounts of good quality datasets, AI can predict the occurrence of numerous natural disasters, which can be the difference between life and death for thousands of people [16]. Some of the natural disasters that can be predicted by AI are:

**Earthquakes:**

AI systems can be trained with the help of seismic data to analyse the magnitude and patterns of earthquakes and predict the location of earthquakes and aftershocks.

**Floods:**

Various researchers and technology experts are developing AI-based applications with the help of rainfall records and flood simulations to predict and monitor flooding.

**Volcanic eruptions:**

AI-powered systems can accurately predict volcanic eruptions with the help of seismic data and geological information.

**Hurricanes:**

AI can use satellite to predict and monitor the path and intensity of hurricanes and tornadoes.

## 5. CONCLUSIONS

The study is assessing new frameworks for effective prevention measures and how AI can fit in and foster the early warning process. So further experiments and understanding the interrelation between AI and big data, what frameworks and systems that worked, and how AI can impact on different business operations whether by introducing new innovations that foster crisis management learning process and early prevention measures. The study from various reviews show promising results in using AI to learn specific industry big data and further evaluation and research is in progress.

## REFERENCES

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, (2015). Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8), pp.228- 232.

[2] A. Gandomi and M. Haider, (2015). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2), pp.137-144.

[3] C. Lynch, (2008). Big data: How do your data grow?, Nature, 455, pp.28-29.

[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, (2015). Significance and challenges of big data research, Big Data Research, 2(2), pp.59-64.

[5] R. Kitchin, (2014). Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1).

[6] C. L. Philip, Q. Chen and C. Y. Zhang, (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275, pp.314-347.

[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, (2014). Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, (2014). On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285, pp.112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, (2014). Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1, pp.114-126.

[10] Xinhua China Daily, (18-Sep-2017). How is AI disrupting financial industry. http://www.chinadaily.com.cn/business/2017-09/18/content_32147126.htm

[11] Focus Elekta's Online Magazine, (2019). How AI is revolutionizing healthcare operations, https://focus.elekta.com/2019/10/how-ai-is-revolutionizing-healthcare-operations/
[12] Louis Columbus, (18-May-2020). 10 Ways AI is improving manufacturing in 2020, Forbes. https://www.forbes.com/sites/louiscolumbus/2020/05/18/10-ways-ai-is-improving- manufacturing-in-2020/?sh=3530e5d1e85a
[13] Kejriwal M. & Zhou P., (2019). SAVIZ: Interactive Exploration and Visualization of Situation Labeling Classifiers over Crisis Social Media Data, International Conference on Advances in Social Networks Analysis and Mining, Vancouver, Aug 27-30, pp705-708.
[14] National Commission, (2011). The Gulf Oil Disaster and the Future of Offshore Drilling.
[15] Dhaimaan Mahmud, (2019). Crisis Management Analysis of the BP Oil Spill, Birmingham Business School.
[16] Naveen Joshi, (15-Mar-2019). How AI can and will predict disasters, Forbes, https://www.forbes.com/sites/cognitiveworld/2019/03/15/how-ai-can-and-will-predict-disasters/?sh=7cddf40d5be2

## AUTHOR

**Prof. Yew Kee Wong** (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, Hewlett-Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# HIGH-FREQUENCY CRYPTOCURRENCY TRADING STRATEGY USING TWEET SENTIMENT ANALYSIS

Zhijun Chen

Department of Financial Engineering, SUSTech University, Shen Zhen, China

## *ABSTRACT*

*Sentiments are extracted from tweets with the hashtag of cryptocurrencies to predict the price and sentiment prediction model generates the parameters for optimization procedure to make decision and re-allocate the portfolio in the further step. Moreover, after the process of prediction, the evaluation, which is conducted with RMSE, MAE and R2, select the KNN and CART model for the prediction of Bitcoin and Ethereum respectively. During the process of portfolio optimization, this project is trying to use predictive prescription to robust the uncertainty and meanwhile take full advantages of auxiliary data such as sentiments. For the outcome of optimization, the portfolio allocation and returns fluctuate acutely as the illustration of figure.*

## *KEYWORDS*

*Cryptocurrency Trading Portfolio, Sentiment Analysis, Machine Learning, Predictive Prescription, Robust Optimization Portfolio.*

## 1. INTRODUCTION

As a decentralized digital asset, cryptocurrency does not exist as physical entity like paper money but secures transaction, controls creation by using strong cryptography [1]. The security and peer-to-peer benefits give bitcoin and many other different types of cryptocurrencies popularity and their markets quickly prosper. Started from 2008, a paper titled “Bitcoin: A Peer-to-Peer Electronic Cash System” marked its inception [2]. By 2017, the price of a single Bitcoin soared 2000% from $863 to $17,550 [3]. Although bubble exists [4], nowadays, the two largest cryptocurrencies, Bitcoin and Ethereum, had a 287.2 billion dollars market capitalization by October 2020, and Bitcoin alone shared 240.6 billion dollars [5].

In this paper, the way of predicting the price of cryptocurrency is by the result of sentiment analysis from tweets which are the short messages in a concise format published in social media platform – Twitter. On average, five hundred million tweets are sent each and every day, and around two hundred billion tweets per year [6]. Such a massive dataset helps this project for the sentiment analysis a lot.

Sentiment analysis or opinion mining combines the usage of natural language processing (NLP), text analysis, computational linguistics, and biometrics to analyse public emotion or preference in the area ranging from marketing to customer survey to recommender system [7][8][9][10]. One of trending task for sentiment analysis is classifying peoples’ emotion into positive, negative or neural to analyse public views towards different topics on social network [11]. In this project,

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 103-113, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111410

sentiment analysis is utilized to generate prediction for the price of cryptocurrency, which then becomes the input of optimization model to work out the capital allocation strategy.

As for the decision, portfolio optimization is another important part in this paper, which is conducted in a prescriptive method. The prescriptions ensure the ability that allocates capital in the robust decision and accommodates the uncertainty in the real world by integrating operations research and management science with machine learning and utilizing both the auxiliary data and the data predicted by machine learning in the process of optimization [12].

## 2. LITERATURE REVIEW

Two main process of this paper – prediction via sentiment analysis and portfolio optimization both have a wide range of related former topics and researches in the financial field.

Emotion affects individual capital allocation strategy according to Amos Tversky and Daniel Kahneman as early as 1979 [13]. After decades of exploration of many great behavioural economists, Paul Tetlock concludes the negative correlation between pessimistic cognition and activeness in stock market [14]. Later, Galen Thomas Panger's research in 'Emotion in Social Media' further bridges the standpoints of behavioural economists and network platform such as social media [15]. Thus, extracting public sentiments for cryptocurrency from tweets in Twitter by data mining and sentiment analysis helps to predict traders' decision and then to predict the price. According to the continuity and time limited of public sentiment, a trading strategy with 2.5 million tweets is established by Hong Kee Sul, Alan R Dennis, and Lingyao Ivy Yuan and produces 11-15% annual returns with a good prospect [16]. Moreover, Y. B. Kim discovers the potential price fluctuation for Bitcoin due to user sentiment [17]. Given by these researches, the topic sentiment analysis for cryptocurrency price prediction is effectively practical.

As for the related works in the field of portfolio optimization, in early 1952, Harry Markowitz formulates the well-known portfolio selection model or mean-variance model [18]. In this model, the portfolio return and risk get measured by expected value and variance respectively. Then, the strategy selection problem is converted into an optimization problem. Based on the key idea of mean-variance model, the Capital Asset Pricing Model (CAPM) is created by William Sharpe in 1964 [19]. These models have a very profound impact on robust portfolio optimization framework, which assumes the worst-casebehaviour faced with unknown parameters or market perturbations [20]. Nowadays, robust portfolio optimization is a wise choice to challenge with parameter uncertainty and estimation errors in portfolio management and to find the optimal portfolio over a basket of cryptocurrencies or the other financial securities with a limited risk.

## 3. RESEARCH METHODOLOGY

In this project, prediction and decision are conducted for Bitcoin and Ethereum, which are two popular cryptocurrencies currently. Six sections comprise the integrity of the trading strategy: Dataset compilation, Sentiment Analysis, Prediction, Measurement, Optimization and Visualization.

### 3.1. Dataset Compilation

Cryptocurrency price dataset or Open High Low Close Volume (OHLCV) is collected from CoinAPI via coinapi-sdk, while tweet dataset collected from tweets with the hash tag of the certain cryptocurrency is compiled from Tweet Archivist started from December twelfth, 2019 to February twelfth, 2020. Coordinate these two types of data set together with respect to the same time interval in hours. That means each row item in the combined data set has both OHLCV and all tweets published in one hour correspondingly.

## 3.2. Sentiment Analysis and Imputation

Sentiment analysis and imputation comprise the further data wrangling.

Firstly, sentiment analysis is to figure out the emotional standpoint of tweets. Tweet text attribute is converted to a bunch of floating-number attributes named ‘pos’, ‘neg’, ‘neu’, and ‘compound’ which respectively represent the positive, negative, neutral viewpoints or emotional tendencies towards certain kind of cryptocurrency. During this process, one general sentiment package nltk.sentiment.vader is utilized. In the package, the model Sentiment Intensity Analyzer is already trained and performs well in daily dialog dataset. A more specialized model for financial or cryptocurrency is regarded to the future work for the project. However, to some extent, the general model fits the cryptocurrency when it comes to casual communication tweets. For example, the sentence “Bitcoin is awesome” has very high compound value, which is 0.6249, while “BAD NEWS FOR #BITCOIN” is totally contrary with a -0.5423-compound value. And a question “I am wondering how people trade on bitcoin” remains neutral, that is 0.0 for compound value.

Next, all the sentiment attributes for tweets are calculated the mean values contributing to the sentiment features for certain hour period which is one of the entries in the final dataset.

Secondly, due to the case that there are some time intervals without tweets for example midnight, it needs imputation for those sentiment attributes to avoid the empty entry. By the assumption of consistency emotion of society, the previous sentiment is pasted to fill the empty row.

Finally, the attribute information of the modified dataset is as follows:

Table 1. Attribute Information for Dataset.

| **Field Name** | **Count** | **Non-Null** | **Data Type** |
|---|---|---|---|
| Period start | 24 | Non-null | Object |
| Period end | 24 | Non-null | Object |
| Time open | 24 | Non-null | Object |
| Time close | 24 | Non-null | Object |
| Price open | 24 | Non-null | Float64 |
| Price close | 24 | Non-null | Float64 |
| Price low | 24 | Non-null | Float64 |
| Price high | 24 | Non-null | Float64 |
| Volume traded | 24 | Non-null | Float64 |
| Trades count | 24 | Non-null | Int64 |
| Change percentage | 24 | Non-null | Float64 |
| Compound | 24 | Non-null | Float64 |
| Pos | 24 | Non-null | Float64 |
| Neg | 24 | Non-null | Float64 |
| Neu | 24 | Non-null | Float64 |

### 3.3. Prediction via Regressor Models

Split the dataset into train set and test set, where regression models are trained and then produce the prediction. During the process of prediction, grid search is used to select the best parameters for predicting the 'Change percentage' target, while the other attributes are the inputs of the prediction models. Four prediction models are used, which are random forest, k-nearest neighbours (KNN), classification and regression tree (CART), and Lasso regressor.

Based on the idea of ensemble learning by training a batch of decision trees [21][22], the pseudo code for random forests or random decision forests is as below. By taking the mode or mean value from the set of trees, the overfitting problem gets released due to avoiding only focusing on individual decision tree model [23].

Algorithm 1: Pseudo code for the random forest algorithm **Error! Reference source not found.**

```
To generate $c$ classifiers:
**for**$i = 1$ to $c$**do**
 Randomly sample the training data $D$ with replacement to produce $D_i$
 Create a root node, $N_i$containing $D_i$
  Call BuildTree($N_i$)
**end for**

**BuildTree(N):**
**if** $N$contains instances of only one class **then**
**return**
**else**
 Randomly select x% of the possible splitting features in $N$
 Select the feature $F$with the highest information gain to split on
 Create f child nodes of $N$,$N_1$, …,$N_f$, where $F$ has$f$possible values ($F_1$, …, $F_f$)
**for**$i = 1$to $f$**do**
        Set the contents of $N_i$to $D_i$, where $D_i$is all instances in $N$that match
                                                         $F_i$
        Call BuildTree($N_i$,)
**end for**
**end if**
```

As a non-parametric method introduced by Thomas Cover in the field of pattern recognition [25], KNN pseudo code is as follow. And the output is regarded as the average of k closet training samples [26].

Algorithm 2: Pseudo code for the KNN algorithm **Error! Reference source not found.**

```
Classify ($\boldsymbol{X}$,$\boldsymbol{Y}$, $x$) // $\boldsymbol{X}$: training data, $\boldsymbol{Y}$: class labels of $\boldsymbol{X}$, $x$: unknown sample
**for**$i = 1$**to**$m$**do**
Compute distance $d(\boldsymbol{X}_i,x)$
**end for**
Compute set $I$containing indices for the $k$smallest distances $d(\boldsymbol{X}_i,x)$
**return** majority label for {$\boldsymbol{Y}_i$where $i \in I$}
```

Due to its precision and explicitness, which are demonstrated in the below pseudo code, decision tree or CART is one of the most widely used machine learning algorithm [28][29].

Algorithm 3: Pseudo code for the CART algorithm **Error! Reference source not found.**

```
d=0, endtree=0
Node(0)=1, Node(1)=0, Node(2)=0
while endtree<1
if Node($2^d$-1) + Node($2^d$) +…+ Node($2^{d+1}$-2) = 2-$2^{d+1}$
      endtree=1
else
      do i=$2^d$-1, $2^d$ ,…,$2^{d+1}$-2
      if Node(i)>-1
      Split tree
      else
        Node(2i+1)=-1
        Node(2i+2)=-1
      end if
      end do
 end if
d = d + 1
end while
```

Aimed to increase the accuracy of the statistical model, lasso regression works both on constructing model by a subset of relevant features and adding information to prevent overfitting and was firstly used in geophysics literature in 1986 [30].

Four machine learning models are used in the project, which are imported from sklearn package. They are Random Forest Regressor, KNeighbors Regressor, Decision Tree Regressor and linear_model.Lasso(). Regressor models, combination with grid searched parameters and cross validated training dataset, get trained by training dataset and then produce the prediction results for test dataset as shown below.

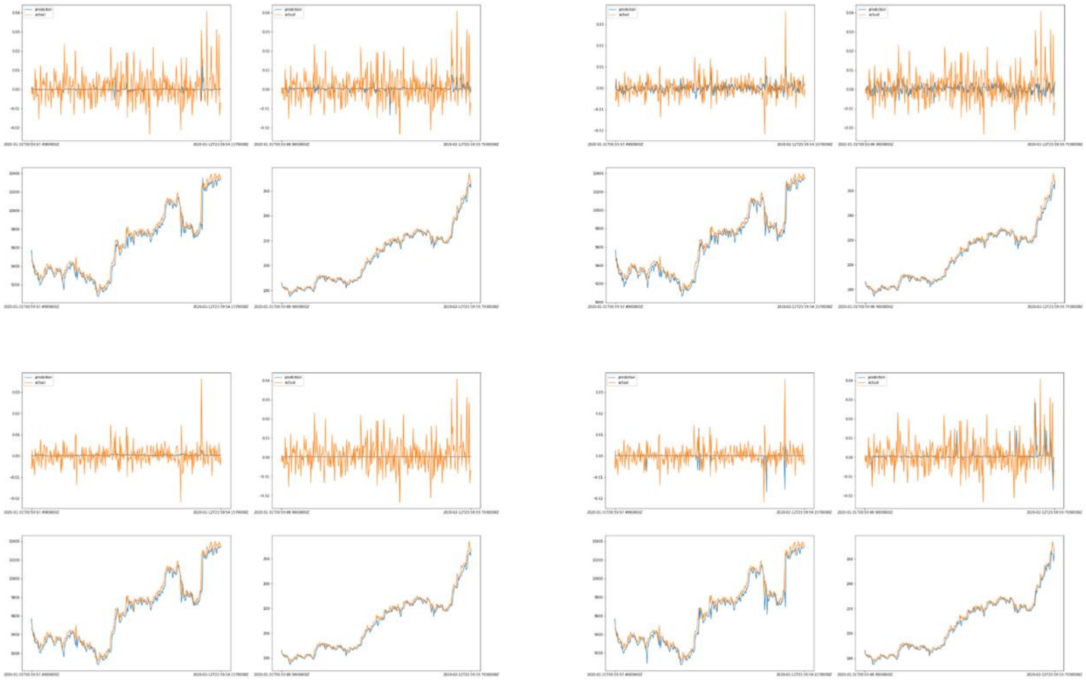


Figure 1. Predictions of change percentage and close price

The sixteen charts in Figure 1 can be divided into four groups. With the writing direction of letter "z", from the position of top left to bottom right, there are the prediction results of random forests, KNN, CART, Lasso regressor respectively. Inside each group, with the same direction, there are the prediction of change percentage for Bitcoin and for Ethereum, the prediction of close price for Bitcoin and for Ethereum respectively. Inside each chart, prediction result, which is the blue line, and actual value, which is the orange one, are plotted in the time interval from January thirty-first, 2020 to February twelfth, 2020.

## 3.4. Measurement for Prediction

To measure the performance of these models and select the most accurate one except the others. Three types of measurements are taken into account, which are root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (denoted by R$^2$).

RMSE is the scale-dependent measurement which calculates the square root of the average of squared errors [31]. And the formula is

$$\text{RMSE}=\sqrt{\frac{\sum_{t=1}^{T}(\widehat{y_t}-y_t)^2}{T}},$$

Where $\widehat{y_t}$ denotes the predicted results for times t,$y_t$ represents the observed values for the same times, and their difference, also called error, is calculated quartic sum over T times [32].

MAE is the arithmetic average for the absolute error and commonly used in time series analysis [34]. With the above denotation, MAE is given by

$$\text{MAE}=\frac{\sum_{t=1}^{T}|\widehat{y_t}-y_t|}{T}\text{Error! Reference source not found.}[34].$$

R$^2$ measures the degree of replication by the proportion of variation [35][36] and is equivalent to the explained sum of squares over the total sum of squares, that is

$$\mathrm{R}^2 = \frac{SS_{reg}}{SS_{tot}},$$
$$SS_{reg} = \sum\nolimits_{t=1}^{T} (\hat{y_t} - y_t)^2,$$
$$SS_{tot} = \sum\nolimits_{t=1}^{T} (y_t - \bar{y})^2,$$

where $SS_{reg}$ is the quadratic sum of the difference between prediction and observed data, and $SS_{tot}$is the squared sum of the difference between observed data and mean value [36].

After comparing the three measurements, RMSE, MAE andR$^2$for four models used to predict the price respectively. Conclusion can be drawn to the case that KNN performs best for the price prediction of Bitcoin while CART predicts the price of Ethereum most accurately, which both have lowest RMSE, MAE and highestR$^2$ (see Table 2 and Table 3)

Table 2. Bitcoin Prediction Measurement.

| | **RMSE** | **MAE** | **R$^2$** |
|---|---|---|---|
| Random forests | 0.008249 | 0.005964 | -0.05151 |
| KNN | 0.008025 | 0.006007 | 0.004652 |
| CART | 0.008293 | 0.005971 | -0.06280 |
| Lasso regression | 0.008070 | 0.005960 | -0.006503 |

Table 3. Ethereum Prediction Measurement.

| | **RMSE** | **MAE** | **R$^2$** |
|---|---|---|---|
| Random forests | 0.007977 | 0.005954 | 0.01649 |
| KNN | 0.008320 | 0.006271 | -0.06975 |
| CART | 0.007717 | 0.005747 | 0.07954 |
| Lasso regression | 0.008105 | 0.005967 | -0.0152 |

As data shows, some of the prediction methods does not perform effectively with respect to their negative R$^2$scores. These may be due to the high volatility of the cryptocurrency market and frequent fluctuation of the cryptocurrency price which decrease the accuracy of prediction.

## 3.5. Portfolio Optimization

Before formulating the strategy, covariance matrix between the history price of Bitcoin and Ethereum needs to be introduced. According to

$$\Sigma = \begin{bmatrix} var(B) & cov(B,E) \\ cov(E,B) & var(E) \end{bmatrix} [37],$$

where $B$ is for the value of Bitcoin, $E$is for the one of Ethereum, $var$ calculates the variance and $cov$ calculates the covariance.

As for the objective function and constraints, according to Bertsimas work in “From Predictive to Prescriptive Analytics”, objective function or predictive prescription is given by

$$\hat{z}_N(x) \in \arg\min_{z \in Z} \sum_{i=1}^{N} w_N^i(x) c(z; y^i) [12],$$

while the constraints – robust maximum return formulation in the worst case are

$$\begin{aligned} &\text{maximize} \min_{\{\mu \in S_m\}} \mathrm{E}[\mathrm{r}_\phi], \\ &\text{subject to} \max_{\{V \in S_v, D \in S_d\}} \mathrm{Var}[\mathrm{r}_\phi] \leq \lambda, \\ &1^T \phi = 1 \ [20]. \end{aligned}$$

In this project, risk λ is taken from the max one of variance of Bitcoin and variance of Ethereum. Transaction cost, which is the extra manipulation cost for re-allocation the capital, is 0.5%. The perception allocation percentage is half and half, that is holding Bitcoin and Ethereum equal amount at the beginning time.

## 3.6. Visualization of Strategy

The results show that all the capital is allocated to buy Bitcoin or Ethereum at every time period as the Figure 2 shows. It seems the capital basket is too monotonous in this case. And the returns for the portfolio tend to soar at some time periods as the illustration of Figure 3.

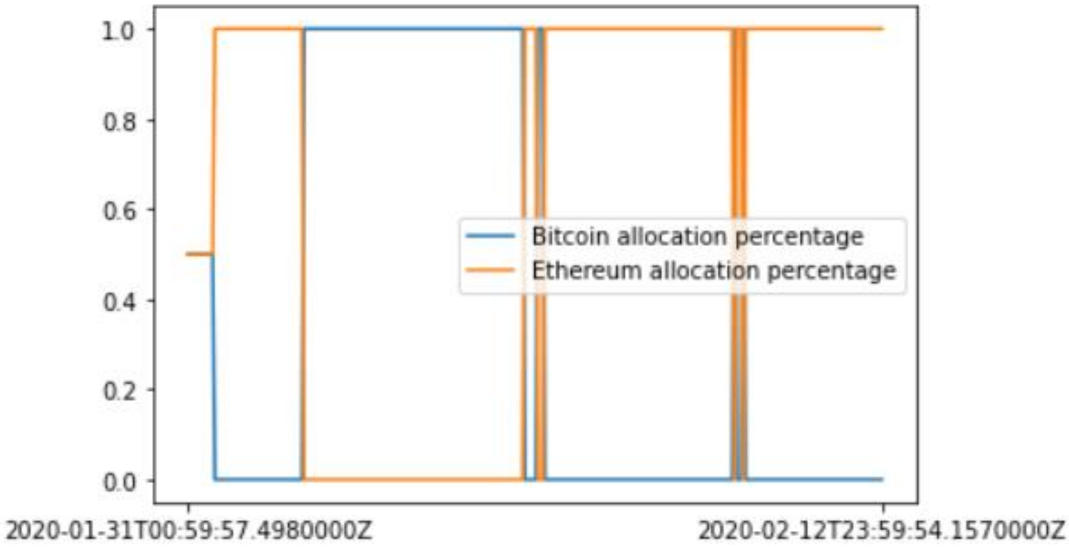


Figure 2. Capital allocation for portfolio

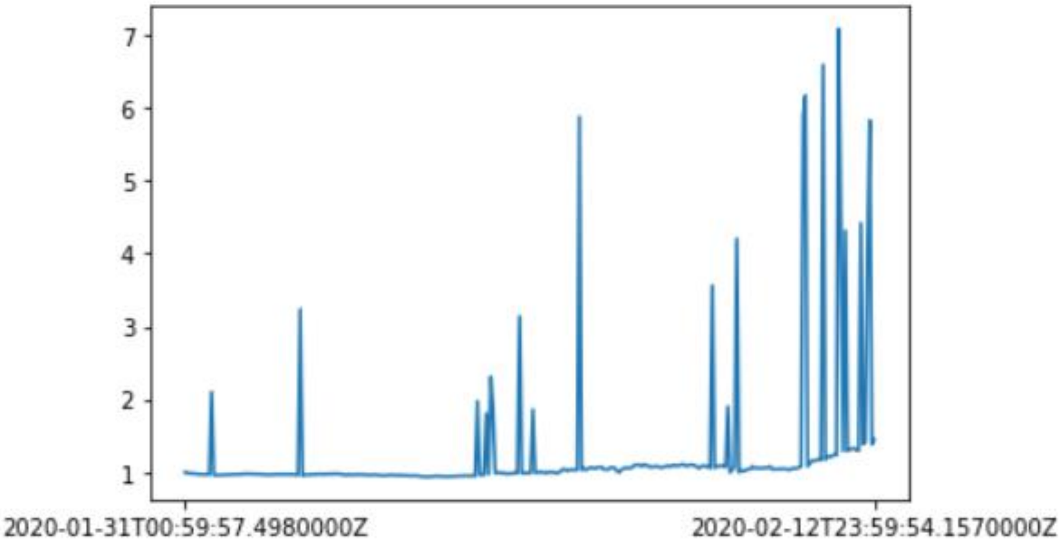


Figure 3. Returns for portfolio

Within the timespan of half month, returns fluctuated acutely with even greater than 7% at some points, but coming down to no more than 1.5% at the end.

## 4. Discussion and Conclusion

The combination of sentiment features and predictive prescription helps to collect auxiliary data from social network and produce an uncertainty-robust portfolio strategy to help capital allocation in the cryptocurrency market.

More work can be done in the future both in the predictive part and prescriptive analysis part, such as replace the general sentiment model by a specific financial-target one, change the risk to a wider range, select more perception allocation percentage, and try different transaction cost, which are all the factors that can have an impact on the results of portfolio construction and final returns of this strategy. Moreover, more trading rules can be included such as short position.

### Acknowledgements

Thanks for my project instructor Hari Bandi and study assistant Roxy Viax, who provided me with comprehensive guides and advices on the process of paper publishment, and Yuxi Li, who always supports me.

### References

[1] Greenberg, A. (2011) "CRYPTO CURRENCY-Money you can't trace". Forbes, 40.

[2] Nakamoto, S. (2008)"Bitcoin: a peer-to-peer electronic cash system". Retrieved from https://bitcoin.org/bitcoin.pdf (accessed April 30, 2018).

[3] Abraham, Jethin & Higdon, Daniel& Nelson, John& and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1 : No. 3 , Article 1. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/1

[4] R. C. Phillips &D. Gorse, (2017) "Predicting cryptocurrency price bubbles using social media data and epidemic modelling", 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7.

[5] "Bit Info Charts" (2013). Available online at: https://bitinfocharts.com (accessed October 24, 2020).

[6] "Internet Live Stats" (2011). Available online at: https://www.internetlivestats.com/twitter-statistics/ (accessed October 24, 2020).

[7] Pang, Bo&Lee, Lillian & Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

[8] Thelwall, Mike& Buckley, Kevan& Paltoglou, Georgios & Cai, Di; Kappas, Arvid (2010). "Sentiment strength detection in short informal text". Journal of the American Society for Information Science and Technology. 61 (12): 2544–2558. CiteSeerX 10.1.1.278.3863. doi:10.1002/asi.21416.

[9] Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. pp. 417–424.

[10] Korkontzelos, Ioannis & Nikfarjam, Azadeh & Shardlow, Matthew & Sarker, Abeed & Ananiadou, Sophia & Gonzalez, Graciela H. (2016). "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts". Journal of Biomedical Informatics. 62: 148–158. doi:10.1016/j.jbi.2016.06.007. PMC 4981644. PMID 27363901.

[11] R. Khan& H. U. Khan& M. S. Faisal&K. Iqbal&M. S. I. Malik, (2016)"An Analysis of Twitter users of Pakistan" Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 8, p. 855.

[12] Bertsimas, D. &Kallus, N., (2014)"From predictive to prescriptive analytics". arXiv preprint arXiv:1402.5481.

[13] Kahneman, D. & Tversky, A. (1979)"Prospect theory: An analysis of decision under risk." Econometrica 47(2), pp263-291.

[14] Tetlock, P.C.(2007)"Giving content to invsotry sentiment: The role of media in the stock market." The Journal of Finance.

[15] Panger, G.T. (2017)"Emotion in Social Media". PhD thesis, University of California, Berkeley.

[16] Hong Kee Sul& Alan R Dennis&Lingyao Ivy Yuan. (2016)"Trading on twitter: Using social media sentiment to predict stock returns". Decision Sciences.

[17] Y. B. Kim& J. Lee, N. Park& J. Choo& J.-H. Kim&C. H. Kim, (2017)"When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation" PLOS ONE, vol. 12, no. 5, p. e0177630.

[18] Harry Markowitz (1952)"Portfolio selection". The Journal of Finance, 7(1):77–91.

[19] Sharpe, W. (1964). "Capital asset prices: A theory of market equilibrium under conditions of risk" J. Finance 19(3) 425–442.

[20] Md. Asadujjaman & Kais Zaman (2014) "Robust Portfolio Optimization under Data Uncertainty" 15th National Statistical Conference, Dhaka, Bangladesh.

[21] Ho, Tin Kam (1995). "Random Decision Forests" (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[22] Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601.

[23] Hastie, Trevor& Tibshirani, Robert& Friedman, Jerome (2008). "The Elements of Statistical Learning (2nd ed.)". Springer. ISBN 0-387-95284-5.

[24] Guo, Hongquan & Nguyen, Hoang & Vu, Diep-Anh & Bui, Xuan-Nam. (2019). "Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach." Resources Policy. 10.1016/j.resourpol.2019.101474.

[25] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879. hdl:1813/31637.

[26] Piryonesi S. Madeh & El-Diraby Tamer E. (2020). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022.

[27] Tay B, Hyun JK, Oh S.(2014)"A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images". Comput Math Methods Med. 2014; 2014:276589. doi: 10.1155/2014/276589. Epub 2014 Jan 21. PMID: 24575150; PMCID: PMC3918356.

[28] Wu, Xindong & Kumar, Vipin& Ross Quinlan, J.& Ghosh, Joydeep & Yang, Qiang & Motoda, Hiroshi & McLachlan, Geoffrey J.& Ng, Angus& Liu, Bing & Yu, Philip S. & Zhou, Zhi-Hua

(2008). "Top 10 algorithms in data mining". Knowledge and Information Systems. 14 (1): 1–37. doi:10.1007/s10115-007-0114-2. ISSN 0219-3116. S2CID 2367747.

[29] Piryonesi S. Madeh & El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.

[30] Santosa, Fadil& Symes, William W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM Journal on Scientific and Statistical Computing. SIAM. 7 (4): 1307–1330. doi:10.1137/0907087.

[31] Hyndman, Rob J.& Koehler, Anne B. (2006). "Another look at measures of forecast accuracy". International Journal of Forecasting. 22 (4): 679–688. CiteSeerX 10.1.1.154.9771. doi:10.1016/j.ijforecast.2006.03.001.

[32] "Coastal Inlets Research Program (CIRP) Wiki - Statistics" (2015). Retrieved 4 February 2015.

[33] Hyndman, R. and Koehler A. (2005). "Another look at measures of forecast accuracy" [1]

[34] Willmott, Cort J.& Matsuura, Kenji (2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Climate Research. 30: 79–82. doi:10.3354/cr030079.

[35] Steel, R. G. D.&Torrie, J. H. (1960). "Principles and Procedures of Statistics with Special Reference to the Biological Sciences". McGraw Hill.

[36] Glantz, Stanton A.& Slinker, B. K. (1990). "Primer of Applied Regression and Analysis of Variance". McGraw-Hill. ISBN 978-0-07-023407-9.

[37] Park, Kun Il (2018). "Fundamentals of Probability and Stochastic Processes with Applications to Communications". Springer. ISBN 978-3-319-68074-3.

## AUTHORS

A student majored in financial engineering in SUS Tech. Trying to undertake the way in Quant for stock market.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# A COMPARATIVE FRAMEWORK FOR EVALUATING CONSENSUS ALGORITHMS FOR BLOCKCHAINS

Dipti Mahamuni

Ira A. Fulton Schools of Engineering – School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

## *ABSTRACT*

*The past five years have seen a significant increase in the popularity of Decentralized Ledgers, commonly referred to as Blockchains. Many new protocols have been launched to cater to various applications serving individual consumers and enterprises. While research is conducted on individual consensus mechanisms and comparison against popular protocols, decision-making and selection between the protocols is still amorphous. This paper proposes a comprehensive comparative framework to evaluate various consensus algorithms. We hope that such a framework will help evaluate current as well as future consensus algorithms objectively for a given use case.*

## *KEYWORDS*

*Consensus Algorithms, Blockchain, Comparative Framework, Decentralized Ledgers.*

## 1. INTRODUCTION

The last five years have seen an unprecedented increase in the number of projects for decentralized ledgers. Primarily, what differentiates one project from another is its core consensus algorithm. What started with the first generation of proof of work systems such as Bitcoin and Ethereum have given way to newer generations of systems that include proof of stake, proof of elapsed time, proof of authority, and other DAG-based consensus systems.

Each of these algorithms features some unique differentiators that are analyzed in literature. Some of these analyses only focus on comparing a handful of algorithms [1], while other studies present the comparative analysis based on a single feature such as performance, security, or cost. For example, Cao et al [2] compare various categories of proof of work, proof of stake, and DAG-based algorithms based solely on their performance characteristics.

This paper presents a comprehensive framework for comparing various consensus algorithms to objectively evaluate various consensus algorithms based on the needs of the project.

The rest of the paper is organized as follows. In section 2, we present the evaluation criteria for comparing consensus algorithms. Section 3 illustrates the decision-making framework and its limitations. Future applications extending this work are presented in Section 4. The paper concludes in the subsequent section.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 115-123, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111411

## 2. EVALUATION CRITERIA

In this section, we present each evaluation criteria for the proposed framework and discuss its importance.

### 2.1. Security

The basic premise of a decentralized ledger is predicated on the security of the consensus algorithm. If an attacker could compromise the security, trust in the entire blockchain can be diminished.

#### 2.1.1. Sybil and Eclipse Attacks

Douceur described the Sybil attack in 2002 [3] where a malicious entity could gain a large influence on the blockchain network by creating a large number of fake identities, devices, IP addresses, or virtual machines. While the Sybil attack tries to subvert the network as a whole, the Eclipse attack tries to prevent an honest node from obtaining the current information by surrounding it with malicious peers.

The current set of consensus algorithms (proof of work, proof of stake, etc) are resistant to Sybil and Eclipse attacks as long as a malicious actor does not get control over a large proportion of hashing power, stake, or the number of DAG nodes.

If this cannot be ensured, then these protocols can fall victim to the Sybil attack [4]. Once the attacker fills the network with malicious clients that are under his control, he possesses control. When all clients work in accordance with the attacker, proof of work is compromised.

A good consensus algorithm must be resistant to these attacks.

#### 2.1.2. 51% attacks

If an attacker can control the majority of the hashing power in a network, then they can control the production of blocks in the network. They can create fraudulent entries in the ledger or prevent legitimate transactions from being recorded in the ledger.

The proof of work algorithm as well as multiple variations of proof of stake algorithms (e.g., basic proof of stake, delegated proof of stake, leased proof of stake, etc.) are particularly vulnerable to the 51% attack.

Sayeed and Marco-Gisbert [5] conclude that in most cases, security techniques usually cannot protect against the 51% attack because the weaknesses are inherited from the consensus protocol. However, a good algorithm must prevent attackers from rewriting history on the blockchain. In the least, the algorithm should make it trivially transparent as to which of the 51+ % of consensus nodes are in agreement while committing fraud.

#### 2.1.3. Internet-based attacks (aka routing attacks, BGP hijacking attacks), DDoS attacks

The attacker uses the internet protocol (IP) and the associated routing protocols such as the Border Gateway Protocol (BGP) to divert traffic away from honest nodes. The attacker can then prevent publication of the generation of honest blocks and push malicious blocks or partition the network to create a double spending attack.

The lack of a trusted messaging system is core to the classic Byzantine General Problem. Even the earliest proof of work algorithm in the Bitcoin network was designed to be a probabilistic solution to this problem as written by Satoshi Nakamoto [6]. However, there have been several successful attacks on these networks over the years.

In addition, a distributed denial of service attack can either take down honest nodes or create artificial partitions in the network.

A good consensus algorithm must show strong resistance to Internet-based attacks. More importantly, the behavior of the network is important when such an attack is underway. While it is acceptable for the network to stop processing transactions when under attack, a good consensus protocol should ensure that an attacker can never compromise the integrity of the ledger. Leaderless DAG-based algorithms perform better than the leader-based algorithms in this regard.

#### 2.1.4. Double Spend attacks

Double spend attacks allow a holder of an asset to spend it more than once. This can happen in multiple situations such as an attacker using a race condition between two transactions before they are finalized on the blockchain (race attack), the merchant not verifying a sufficient number of blocks for finality (Finney attack), or the attacker creating a fraudulent block that is not confirmed on the blockchain.

Both proof of work and proof of stake algorithms present vulnerabilities for double spend. In a slow proof of work system, an attacker can use the time required to create multiple blocks to launch a double spend attack. The proof of stake algorithms are vulnerable to double spending attacks due to a problem called “nothing at stake” [7]. This means that if a malicious node has nothing in its stake, then it has nothing to lose and nothing to counteract its malicious actions.
A good consensus algorithm should demonstrably prevent double spend attacks.

### 2.2. Decentralization

Decentralization is critical to the operation of any blockchain. Without decentralization, the blockchain degenerates to a centralized ledger or a database and is therefore susceptible to manipulation by a single or a small number of organizations.

The selection of an algorithm directly influences the degree of decentralization. For example, proof of stake’s miner selection is done based on the assets, or amount of cryptocurrency, that a miner owns. Because of this, the algorithm is prone to becoming centralized over time, allowing richer accounts (known as whales) to have more control over the blockchain [7].

The delegated algorithms (such as Delegated Proof of Stake or Delegated Byzantine Fault Tolerant algorithms) typically elect delegates who must reach consensus to verify and add a block to the blockchain [7]. Since there are a limited number of delegates, there is a risk of the system becoming centralized.

Let us look at other factors that affect the decentralization of a consensus-based network.

### 2.2.1. Decentralization through scale

The scalability of the nodes that run the ledger is important to the network. The trustworthiness of a network increases substantially as the number of nodes that run the network increases. However, this requires extra communication and additional time to reach consensus.

A good consensus algorithm must be able to scale to a large number of nodes. Typically, there are tradeoffs between the extra security and trust added by scaling to a large number of nodes and the communication and latency overheads brought on by large-scale networks.

### 2.2.2. Geographical distribution of nodes

Though this is not necessarily a consensus protocol design issue, if any aspect of the consensus protocol encourages the concentration of nodes in one country or a geographic area, then the trustworthiness of that ledger is reduced.

For example, a disproportionate number of Bitcoin blocks are mined out of a few countries due to the relatively cheap electric power required for power-hungry proof of work algorithms. A good protocol should prevent such geographical concentration from occurring.

### 2.2.3. Permissioned versus Permissionless

A good consensus protocol should be able to run in a permissionless environment. It should be possible for anybody in the world to validate the transactions on a blockchain without explicitly asking permission from the existing set of nodes.

### 2.2.4. Open source - decentralization of development and source code

The consensus algorithm as well as the source code should be open source. Any security vulnerabilities can be easily found by the open source community. Similarly, this eliminates over-dependence on a few smart engineers for continued enhancement of the algorithm.

### 2.2.5. Requirement for a specialized hardware

Some consensus algorithms require specialized hardware to function effectively. For example, consensus algorithms based on Proof of Elapsed Time [7] rely on specialized hardware present on Intel CPUs (SGX) supporting Trusted Execution Environment (TEE). Similarly, a lot of Ethereum mining is accelerated using specialized GPUs today.

A good consensus algorithm should ensure that it does not need any specialized hardware. More importantly, it should ensure that the presence of specialized hardware, including but not limited to GPUs, FPGAs, and ASICs, should not provide any unfair advantage in creating blocks on the blockchain.

## 2.3. Scalability

### 2.3.1. Energy Consumption

The proof of work algorithm requires a lot of computational power for a miner to be able to add a block to the blockchain. This computation uses an excessive amount of electricity as compared to proof of stake [4]. The proof of stake algorithm reduces computational power, hence reducing energy consumption. Panda et al [4] also conclude that the proof of burn algorithm has a better energy consumption rate than proof of work does.

A good consensus algorithm should process transactions with minimum electricity consumption.

#### 2.3.2. Finality

The time to finality of a transaction is defined as the elapsed time from when a transaction is submitted to the network to the time the transaction is recorded in the blockchain. Traditional proof of work algorithms lack a strict definition of finality since waiting for more blocks in the blockchain only increases the probability of the transaction being final. Chaudhry & Yousef [8] present a table that lists probabilistic versus deterministic finality times of various algorithms. Since the proof of stake algorithm spends less time doing complex computations, block finality time is faster than it is for proof of work [4]. Newer DAG-based consensus algorithms have further reduced this time to a few seconds.

Many practical applications, such as credit card transaction processing, require fast finality times. A good consensus algorithm should support finality in a few seconds.

#### 2.3.3. Throughput (Transactions per Second)

Large networks that have high transactions per second cannot use the proof of work algorithm due to its time-consuming nature [7]. Using proof of work, solving the hash puzzle is a difficult and time-consuming task. Without the right hardware, solving the hash could take even longer. This reduces the transaction processing rate, or the throughput.

Since the proof of stake algorithm spends less time doing complex computations, it can process far more transactions every second.

Fast finality times often go hand in hand with high throughput requirements. For example, processing credit card transactions also requires high throughput.

### 2.4. Governance

#### 2.4.1. Fork Resistance

Forks are bad in a consensus network. They essentially create multiple sources of truths that counter the establishment of trust in transactions recorded on a blockchain. Forks also create opportunities for double-spend attacks since a different version of the truth can be recorded in each fork.

Some consensus algorithms cannot avoid forking, at least temporarily, due to network latency and miner behavior. Neudecker and Hartenstein [9] have empirically analyzed forking in the Bitcoin network and concluded that the probability of a block to become part of the main chain increases linearly from its creation. This time window creates an opportunity for double spend.

While the issues created by this type of temporary forks can be mitigated by waiting for a sufficiently large number of subsequent blocks, it is the hard forks that cause major problems in the user community. Hard forks can rewrite the blockchain and make previously valid blocks invalid, or vice versa.

A good consensus protocol should have built-in mechanisms to detect as well as to deter hard forks.

#### 2.4.2. Software Upgrades

Another aspect of the governance of a consensus algorithm is how the algorithm itself is updated. The algorithm needs to be updated from time to time to allow for critical security-related changes, fixing software bugs, adding more features, or for performance improvements.

Ensuring that the consensus algorithm can be leveraged for these critical software upgrade decisions is important for the viability of that blockchain.

### 2.5. Compliance

The first generation of consensus protocols was perceived as a way to get around the governmental regulatory and compliance issues by providing pseudonymity and confidentiality. Today, as governments across the world start looking into regulations for blockchain usage, it is important that the next generation of consensus protocols look at compliance-related features of the blockchain.

#### 2.5.1. Regulatory compliance (e.g., KYC and AML)

Legal financial applications of the blockchain will subject the blockchain networks to a similar level of scrutiny as any other financial institution. Two primary regulatory requirements are likely to become important here.

Know Your Customer (KYC) requires financial institutions to verify certain aspects of a user. The protocol should ensure that it can exert some control over user accounts to allow for this.

Anti-Money Laundering (AML) and other fraud detection systems require the ability to monitor all transactions on the network and query history of transactions. Consensus protocols should have features to transparently distribute such information.

#### 2.5.2. Removal of illegal content

Blockchains are built on the premise of immutability, and therefore it becomes impossible for blockchains to remove content. There are several instances of storage of illegal content (such as stolen classified documents) on public blockchains like the Bitcoin network. Short of a hard fork, removing this content violates the basic principle of immutability.

A good protocol should have a legitimate mechanism for the nodes to reach consensus to alter the blockchain in a controlled and transparent manner.

## 3. THE FRAMEWORK

With the backdrop of the discussion in the section above, Table 1 presents a framework to evaluate the consensus algorithms. The methodology used in determining the framework is as follows. The primary reason for using a blockchain in a business application is the trust and security that is ensured by decentralization. Hence, this model assigns over half of the weightage to security and decentralization requirements. However, we have seen that many initial deployments of blockchain applications based on the early proof of work systems have suffered due to the inherent lack of scalability or governance models. Further, some governments have issued bans on certain blockchains due to lack of regulatory compliance requirements. This model emphasizes the importance of these requirements in today's blockchain systems by assigning nearly half the weightage to considerations related to scalability, governance and compliance. By balancing the security and decentralization requirements with those related to scalability, governance and compliance, this model achieves a holistic evaluation framework.

The weights provided in Table 1 are based on the analysis of the most common use cases of blockchain today. Based on the application that runs on the blockchain, the user can assign appropriate weights to each of the evaluation criteria. For some applications, scalability and time for finality may be much more important than resistance to double spending attacks (for example, it is inconceivable today to run all credit card transactions on the Bitcoin network because a user in front of a gas station will be unwilling to wait for several minutes for the block to be confirmed), while for other applications, security features might be far more important than the speed (for example, applications that register a deed for the ownership of a house on the blockchain). Based on a given application, the framework can be used to update the weights, and then a weighted score can be calculated to determine the best consensus algorithm for the given application.

Table 1. Comparative Analysis Framework

| Evaluation criteria | Description | Weightage |
|---|---|---|
| Security | | 25% |
| | Resistance to Sybil and Eclipse attacks | 7% |
| | Resistance to 51% attack | 7% |
| | Resistance to Internet-based attacks | 4% |
| | Resistance to Double Spend attacks | 7% |
| Decentralization | | 30% |
| | Decentralization through scale | 7% |
| | Geographical distribution of nodes | 7% |
| | Permissionless | 3% |
| | Open source | 7% |
| | Non requirement for specialized hardware | 6% |
| Scalability | | 20% |
| | Energy consumption | 7% |
| | Finality | 6% |
| | Throughput (Transactions per Second) | 7% |
| Governance | | 15% |
| | Fork resistance | 8% |
| | Software upgrades | 7% |
| Compliance | | 10% |
| | Regulatory compliance | 7% |
| | Removal of illegal content | 3% |

## 3.1. Limitations

As discussed above, the model presented above is generic in nature. While it is useful to select a blockchain consensus protocol for the vast majority of business and consumer applications, the model does not work for every application. There might be superseding considerations or extenuating circumstances such as geographical data sovereignty requirements, or integrations with existing infrastructure that influence the decision of the consensus protocol. In such cases, the model presented above should be used in conjunction with other considerations.

## 4. FUTURE WORK

This paper presents a single evaluation framework, and as discussed in section 3.1, the presented framework may not be suitable for some niche use cases. The creation of multiple bespoke frameworks designed for a set of specific applications is recommended to address the needs of these use cases.

Currently, the framework is presented as a tool where the decision maker will enter scores against each criterion. We recommend the creation of an automated test-suite that can run against a target blockchain to evaluate and generate an automated score. An open source project that implements the test-suite could pave the way for objectively measuring effectiveness of the consensus protocols. Such a test tool can also inspire the design of a “perfect blockchain protocol” that can objectively maximize the score relative to the decision-making criteria described above.

## 5. CONCLUSIONS

In this paper, we discussed objective criteria for evaluating various consensus algorithms of a blockchain network. These criteria range from Security and Decentralization features, Scalability and Cost, Governance, and Compliance. We then presented a decision-making framework that holistically balances these criteria for a vast majority of business and consumer applications. This model can be used to make objective decisions about the selection of a consensus algorithm for blockchain based projects, as well as the comparison of new consensus algorithms against the existing ones.

## REFERENCES

[1] Ashar Ahmad, Abdulrahman Alabduljabbar, Muhammad Saad, DaeHun Nyang, Joongheon Kim, & David Mohaisen, (2021) “Empirically comparing the performance of blockchain's consensus algorithms”, IET Blockchain Vol. 1, No. 1, pp. 56-64.

[2] Bin Cao, Zhenghui Zhang, Daquan Feng, Shengli Zhang, Lei Zhang, Mugen Peng, Yun Li (2020) “Performance analysis and comparison of PoW, PoS and DAG based blockchains”, Digital Communications and Networks, Vol. 6, Issue 4, Nov 2020, pp480-485.

[3] Douceur, J.R. (2002) “The Sybil Attack”, Revised Papers from the First International Workshop on Peer-to-Peer Systems, Springer: London, UK, pp. 251–260.

[4] S. S. Panda, B. K. Mohanta, U. Satapathy, D. Jena, D. Gountia and T. K. Patra, (2019) "Study of Blockchain Based Decentralized Consensus Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), pp. 908-913.

[5] Sarwar Sayeed and Hector Marco-Gisbert (2019) “Assessing Blockchain Consensus and Security Mechanisms against the 51% Attack”, Applied Sciences, Vol. 9, Issue 9, Nov 2020, pp. 1788.

[6] Satoshi Nakamoto (2008) “Re: Bitcoin P2P e-cash paper”, Available online at: https://www.mail-archive.com/cryptography@metzdowd.com/msg09997.html

[7] Seyed Bamakan, Amirhossein Motavali, & Alireza Bondarti, (2020) “A survey of blockchain consensus algorithms performance evaluation criteria”, Expert Systems with Applications Vol. 154, pp. 113385.

[8] N. Chaudhry and M. M. Yousaf, (2018) "Consensus Algorithms in Blockchain: Comparative Analysis, Challenges and Opportunities," 2018 12th International Conference on Open Source Systems and Technologies (ICOSST), pp. 54-63

[9] Till Neudecker and Hannes Hartenstein. (2019) “Short Paper: An Empirical Analysis of Blockchain Forks in Bitcoin”, Lecture Notes in Computer Science book series (LNCS), Vol 11598, pp. 84-92.

## AUTHORS

Dipti Mahamuni is a senior year student at Arizona State University pursuing her BS in Computer Science from the Ira A. Fulton Schools of Engineering - School of Computing, Informatics, and Decision Systems Engineering.
Her areas of interest include Blockchain, AI/Machine learning and IoT.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# DMC: DECENTRALIZED MIXER WITH CHANNEL FOR TRANSACTION PRIVACY PROTECTION ON ETHEREUM

Su Liu and Jian Wang

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

## ABSTRACT

*Ethereum is a public blockchain platform with smart contract. However, it has transaction privacy issues due to the openness of the underlying ledger. Decentralized mixing schemes are presented to hide transaction relationship and transferred amount, but suffer from high transaction cost and long transaction latency. To overcome the two challenges, we propose the idea of batch accounting, adopting batch processing at the time of accounting. For further realization, we introduce payment channel technology into decentralized mixer. Since intermediate transactions between two parties do not need network consensus, our scheme can reduce both transaction cost and transaction latency. Moreover, we provide informal definitions and proofs of our scheme's security. Finally, our scheme is implemented based on zk-SNARKs and Ganache, and experimental results show that the higher number of transactions in batch, the better our scheme performs.*

## KEYWORDS

*Ethereum, transaction privacy, decentralized coin mixer, payment channel, zero-knowledge proof.*

## 1. INTRODUCTION

During the past few years, the blockchain technology has drawn tremendous interests from IT industries (e.g. Google, Alibaba and Amazon) to financial institutions (e.g., Goldman Sachs and JP Morgan). Currently, main application areas of blockchain involve finance, payment, data services and so on. Especially in the financial industry, transaction is a significant component, representing the main financial activity of enterprises and individuals. The importance of data places great demands on security and privacy of blockchain.

Ethereum is a distributed append-only public transaction ledger maintained by consensus protocols. However, it suffers from transaction privacy leakage [1], [2] due to the decentralized nature of blockchain. Though the generated account addresses are pseudonymous, but it is possible to link these addresses with real world identities by deanonymization techniques, such as address clustering [3], [4] and transaction graph analysis [5]. Furthermore, transaction relationship and transferred amount between users can be directly obtained via analysing the underlying public ledger, and moreover, attackers can infer users' income levels, spending habits, etc. Therefore, the openness and sensitiveness of transaction information force Ethereum community to design solutions to guarantee transaction privacy.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 125-145, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111412

Recently, some schemes and projects [6]-[9], [11], [13]-[15] have been proposed, attempting to solve the privacy problem while ensuring the verifiability of transactions and the reliability of the ledger. Among them, the main method is coin mixer [8], [9], [11], [13]-[15], where senders deposit some coins into a centralized third party or a smart contract, then the third party or contract transfers the equivalent coins to receivers when they withdraw from the mixer. These mixing mechanisms can be categorized in two classes: (i) centralized mixer, simple but lack of security and privacy; (ii) decentralized mixer, secure but heavy computation. In the following, we will focus on the problem of decentralized mixer.

In the decentralized mixer with any mixing amount, compared with fixed mixing amount, there are three main operations: (i) deposit, the sender deposits some coins into mixer contract in the form of note; (ii) transfer, the sender exploits created notes inside the contract to transfer to the recipient; (iii) withdraw, the recipient redeems the corresponding coins from the contract. Notably, the verification of transactions generated by the above three operations will consume hundreds of thousands gas for heavy cryptographic computation. Moreover, the high gas price (about $\$24/10^6$ gas at the time of writing) due to expensive computing resources will impose a further cost burden. On the other hand, the transaction latency due to the underlying mechanism (block time interval of about 15 seconds and transaction confirmation time of about 10 minutes [16]) and network congestion is also intolerable. In all, the costly transaction fee (equal to the multiplication of gas used and gas price) and the long transaction latency make it challenge to put the decentralized mixing scheme into practice.

To address the aforementioned issues, we introduce the concept of batch accounting. It means that not every time a transaction occurs, it is submitted to the blockchain, but when several transactions are completed, the final transaction result is recorded on the blockchain. Inspired by payment channel technique, whose key idea is to only record the final transaction result on the blockchain, ignoring intermediate transaction process between two parties, we employ payment channel as the technical support behind batch accounting, proposing a decentralized mixer with channel (DMC). The original transfer operation in the above mixing scheme is completed through payment channel, i.e., off-chain transmission of transaction messages. More specifically, when need to transact, the sender utilizes notes in mixer contract to create a channel note, served as a payment channel. The deposit in the channel is equal to the denomination of the channel note. Then via the channel, the sender creates new transactions (including the total amount that the sender needs to transfer to the recipient so far) and sends them directly to the recipient through anonymous network. After receiving a transaction, the recipient verifies and decides whether to accept it. Essentially, the transaction between two parties is the redistribution of the deposit in the channel. When not need the channel again, the recipient closes the channel by posting the latest received transaction to blockchain.

Since intermediate transactions between two parties do not go through the Ethereum network, no decentralized consensus is required. Hence, our scheme can achieve the reduction in transaction cost and latency. Specifically, due to being free from the influence of underlying mechanism and network congestion, the latency of off-chain transactions can be decreased to the communication delay of anonymous network. On the other hand, although the recipient still needs to verify the correctness of off-chain transactions, the verification is performed locally rather than on the Ethereum virtual machine (EVM), so our scheme gets rid of the expensive computing resources on Ethereum and achieves very low transaction cost. Compared with original transactions, the cost and latency of off-chain transactions are both negligible.

**Contributions.** In summary, the main contributions of this paper are as follows.

(1) We propose the idea of batch accounting and construct a decentralized mixer with channel called DMC, which reduces transaction cost and latency while hides the transaction relationship and transferred amount. Our scheme is suitable for frequent transactions between senders and recipients. Because there is no need for network consensus for most transactions of two parties, the reduction in cost and latency is achieved.

(2) We implement DMC based on zk-SNARKs and Ganache, and conduct a number of experiments to evaluate its performances. The results show that our scheme is feasible. Combined with theoretical and experimental analysis of DMC, we can get the average transaction cost and transaction latency are both approximately $1/n$ of other mixing schemes, $n$ denoting the unfixed number of transactions in batch.

**Paper Organization.** The rest of the paper is organized as follows. Section 2 provides some background on Ethereum and payment channel, and explains the cryptographic primitives. Section 3 describes the decentralized mixing scheme. Then, we construct our scheme DMC based on the above decentralized scheme in Section 4. Furthermore, Section 5 details the implementation of DMC, evaluates its performance, and draws comparisons with other schemes. The related work is reviewed in Section 6. Finally, we conclude this paper in Section 7.

## 2. PRELIMINARIES

In this section, we outline the related background of Ethereum and payment channel. In addition, we describe the cryptographic primitives for DMC: commitment scheme, public key encryption scheme and zero-knowledge proof zk-SNARKs.

### 2.1. Ethereum

In Ethereum, account is an important concept, indexed by address. There are two types of accounts — Externally Owned Account (EOA), representing a user with a pair of public key $pk$ and secret key $sk$; and contract account, representing a smart contract with code and storage. The interaction between accounts is made through transactions generated by EOA. A transaction is composed of the destination account address, the transferred amount $v$, an optional data field $data$ (specifying the called function and the passed parameters), and a signature, etc. In this paper, we denote a specific transaction by $\mathtt{tx} = (data)$.; if the transferred amount $v$ exists, it should be pointed out in addition. Each transaction needs to pay a certain transaction fee for operations made in the transaction, which uses gas as the unit for measuring the computational and storage resources. Take some contract operations for example, storing costs 20,000 units of gas while writing and reading cost 5,000 and 200 respectively [18].

### 2.2. Payment Channel

The payment channel technology [19], [20] is an important proposal to address the challenges of the scalability and transaction fee. Lightning network [21] and Raiden network [22] are popular examples deployed on Bitcoin and Ethereum respectively. In Bitcoin or Ethereum, a payment channel is corresponding to a multi-signature address or a smart contract. The payment channel technique includes three procedures: (i) opening a channel, the sender deposits into a multi-signature address/smart contract to create a payment channel; (ii) off-chain transactions, the sender sends signed transaction messages directly to the recipient without passing through the blockchain network; 3) closing a channel, the recipient withdraws from the channel and the remaining coins are returned to the sender. Essentially, the transactions between the two parties are the redistribution of the deposit in the channel.

## 2.3. Cryptographic Building Blocks

Next, we will describe the cryptographic building blocks involved in our scheme. More details about these cryptographic primitives are available in [23]. In the following, λdenotes the security parameter and CRH denotes collision-resistant hash function.

**Commitment Scheme.** A commitment scheme is composed of two algorithms (Comm, Open) such that:

•$cm$ ← Comm $(m, r)$: given message $m$ and randomness $r$, output commitment $cm$.
• $\{0, 1\}$ ← Open $(cm, m, r)$: given commitment $cm$, message $m$ and randomness $r$, output 1 if $cm$= Comm$(m, r)$ and 0 otherwise.

For the purposes of this paper, we will use the commitment scheme which is statistically binding and computationally hiding.

**Public Key Encryption Scheme.** A public key encryption scheme comprises a triple of algorithms ( KeyGen, Encrypt, Decrypt) such that:

•$(pk, sk)$ ← KeyGen($1^{\lambda}$): given security parameter $1^{\lambda}$, output a pair of public key $pk$ and secret key $sk$.
•$c$ ← Encrypt $(pk, m)$: given public key $pk$ and plaintext $m$, output ciphertext $c$.
•$m$ ← Decrypt $(sk, c)$: given secret key $sk$ and ciphertext $c$, output plaintext $m$ or an invalid symbol ⊥.

In this paper, we directly adopt user's public and secret key on Ethereum. So, it's a good choice to use *eciespy* [24], an Elliptic Curve Integrated Encryption Scheme for Ethereum.

**Non-Interactive Zero-Knowledge Proof (NIZK).** A non-interactive zero-knowledge proof is a two-party protocol between a prover and a verifier with two stages. At the proving stage, the prover uses private data to generate the proof without interaction with the verifier. At the verifying stage, the verifier checks the validity of the proof while obtaining no more information.

Let $\mathcal{R}$ be a binary relation for instance $x$ and witness $\omega$, and let $\mathcal{L}$ be corresponding language $\mathcal{L} = \{x \mid \exists\omega\colon (x, \omega) \in \mathscr{R}\}$. NIZK is a protocol where a prover tries to convince a verifier that an instance $x$ is in the language $\mathcal{L}$. In addition, NIZK can permit proving computational statements, but the computational problem needs to be converted into an arithmetic circuit.

A NIZK for arithmetic circuit $\mathcal{C}$ consists of four algorithms (Setup, KeyGen, Prove, Verify) such that:

• $(pp)$ ← $\mathrm{Setup}_{\mathrm{zkp}}(1^{\lambda})$: given security parameter λ, output public parameters $pp$.
• $(pk, vk)$ ← $\mathrm{KeyGen}_{\mathrm{zkp}}(pp, \mathcal{C})$: given $pp$ and arithmetic circuit $\mathcal{C}$, output proving key $pk$ and verification key $vk$.
• $\pi$ ← Prove $(pk, x, \omega)$: given proving key $pk$, instance $x$ and witness $\omega$, output non-interactive proof $\pi$ if $(x, \omega) \in \mathcal{R}$.
• (0, 1) ← Verify $(vk, x, \pi)$: given verification key $vk$, instance $x$, and a proof $\pi$, output 1 if $x \in \mathscr{L}$; otherwise 0.

In this paper, we use zero-knowledge Succint Non-interactive ARgument of Knowledge (zk-SNARK), the most preferable NIZK that has succinct proof size and sublinear verification time. zk-SNARK satisfies the following properties: completeness, succinctness, soundness and zero-knowledge.

## 3. DECENTRALIZED MIXER

In this section, we present a general decentralized mixing scheme by summarizing existing mixing schemes. It utilizes commitment scheme to hide the transaction information and meanwhile applies zero-knowledge proof to ensure the validity of transactions. We firstly provide data structures used in decentralized mixer, then describe its mixing mechanism.

### 3.1. Data Structures

**Note.** Like a cheque, a note $note$ includes an owner $pk$, a denomination $v$ and a random number $r$ (to ensure the uniqueness), i.e., $note = (pk, v, r)$. The following two concepts are associated to a note.

- Commitment, $cm$ = Comm $(pk, v, r)$: obviously, a commitment $cm$ is the commitment to a note $note = (pk, v, r)$, used to hide specific information of the note.
- Serial number, $sn$ = CRH $(sk, r)$: a serial number $sn$, also called nullifier, is the hash of $r$ and the secret key $sk$ corresponding to $pk$, used to prevent double-spending issues.

**CMTree.** CMTree denotes a Merkle tree whose leaf nodes are commitments of created notes. The existence of commitments in CMTree are viewed as proof of ownership of coins in the mixer. Every time a commitment is inserted, the root of CMTree is updated. These generated Merkle roots are then added to an array, denoted by Roots.

**SNSet.** To prevent double-spending issues, all serial numbers of spent notes are recorded in an array, denoted by SNSet. If the corresponding serial number is in SNSet, it indicates that the note has been spent, otherwise the note can be spent.

### 3.2. The Mixing Mechanism

In Ethereum, the decentralized mixer is implemented by smart contract. Users make interactions with the mixer contract to deposit, transfer and withdraw. The decentralized mixing mechanism, described in Figure 1, consists of the next three components. Note that we ignore the Create Account algorithm, because it is the same as the creation of accounts in Ethereum. The original accounts in Ethereum are completely compatible with our scheme.

**Setup.** The setup algorithm is executed only once by a trusted third party (TTP) to generate public parameters and to deploy a mixer contract. Note that the setup algorithm can use secure multi-party computation techniques to mitigate the trust requirement for TTP.

**User Algorithms.** A user can run the following algorithms to interact with the mixer contract and create valid transactions. For convenience, take sender Alice (A) and recipient Bob (B) for instance.

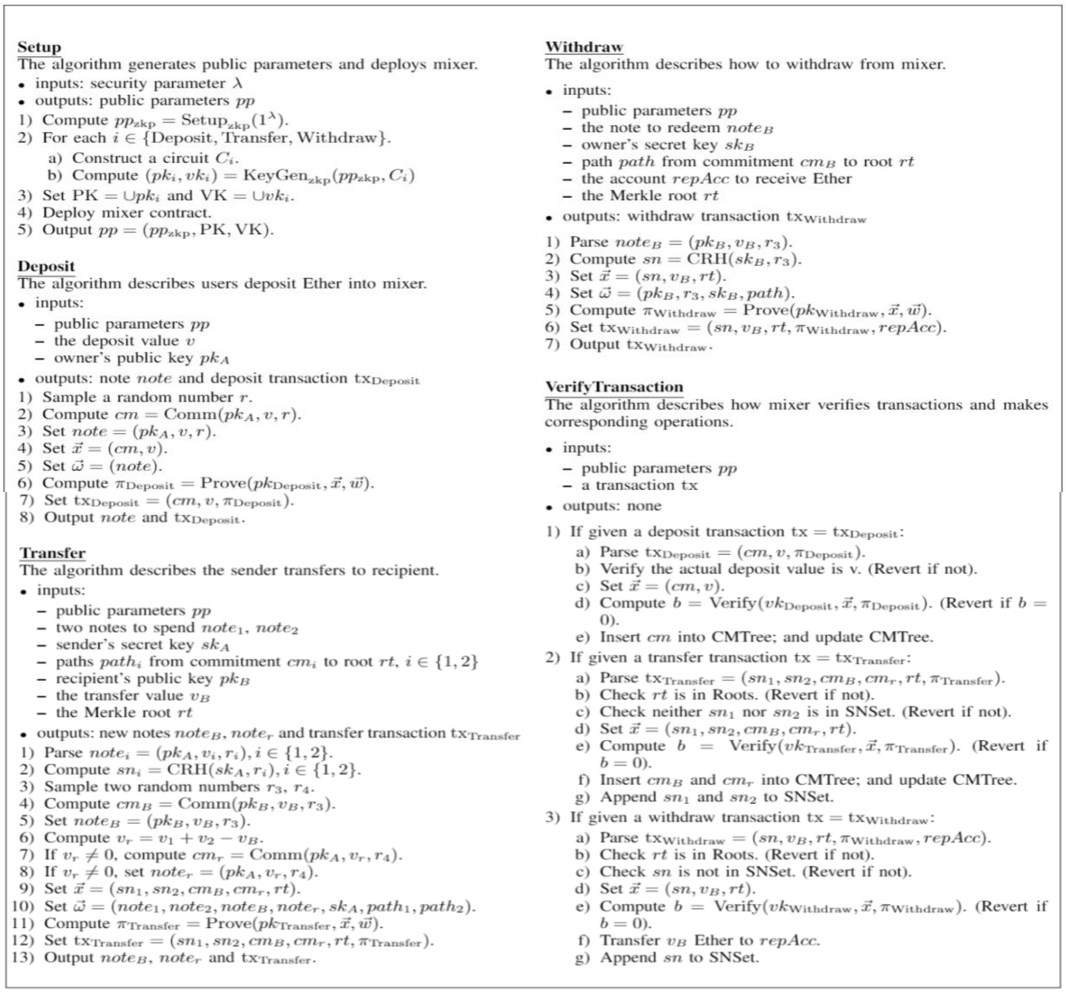
- Deposit: The Deposit algorithm is to convert some Ether into an equivalent note, e.g., Alice deposits $v$ Ether to mixer.
- Transfer: The Transfer algorithm is to destroy some old notes and create some new notes. For example, Alice uses her two notes to transfer $v_B$ Ether to Bob.

• Withdraw: The Withdraw algorithm is to redeem some note into equivalent Ether, e.g., Bob redeems $v_B$ Ether from mixer.

**Mixer Contract.** Firstly, users use one of three algorithms mentioned above to generate corresponding transactions, and send them to the blockchain network. After users submitting transactions, the mixer contract verifies and conducts related operations according to the Verify Transaction algorithm. Specifically, the mixer contract is responsible for verifying the correctness of transactions (e.g., verifying zero-knowledge proofs and serial numbers), and if passing the verification, making corresponding changes (e.g., inserting new commitments into CMTree and serial numbers into SNSet).

## 4. DMC: DECENTRALIZED MIXER WITH CHANNEL

The section gives a detailed description of our scheme based on the decentralized mixing scheme in Section 3. We first discuss the intuition of the scheme based on the next three attempts, then describe the specific construction of the scheme Π = (Setup, Deposit, Open-Channel, Offchain-Transfer, CloseChannel, Withdraw, VerifyTransaction, VerifyOffchainTransfer). At last, we provide the security definitions and proofs of our scheme.



**Setup**

The algorithm generates public parameters and deploys mixer.

```
• inputs: security parameter $\lambda$
• outputs: public parameters $pp$
1) Compute $pp_{\text{zkp}} = \text{Setup}_{\text{zkp}}(1^\lambda)$.
2) For each $i \in \{\text{Deposit, Transfer, Withdraw}\}$.
    a) Construct a circuit $C_i$.
    b) Compute $(pk_i, vk_i) = \text{KeyGen}_{\text{zkp}}(pp_{\text{zkp}}, C_i)$
3) Set $\text{PK} = \cup pk_i$ and $\text{VK} = \cup vk_i$.
4) Deploy mixer contract.
5) Output $pp = (pp_{\text{zkp}}, \text{PK}, \text{VK})$.
```

**Deposit**

The algorithm describes users deposit Ether into mixer.

```
• inputs:
  – public parameters $pp$
  – the deposit value $v$
  – owner's public key $pk_A$
• outputs: note $note$ and deposit transaction $\text{tx}_{\text{Deposit}}$
1) Sample a random number $r$.
2) Compute $cm = \text{Comm}(pk_A, v, r)$.
3) Set $note = (pk_A, v, r)$.
4) Set $\vec{x} = (cm, v)$.
5) Set $\vec{\omega} = (note)$.
6) Compute $\pi_{\text{Deposit}} = \text{Prove}(pk_{\text{Deposit}}, \vec{x}, \vec{w})$.
7) Set $\text{tx}_{\text{Deposit}} = (cm, v, \pi_{\text{Deposit}})$.
8) Output $note$ and $\text{tx}_{\text{Deposit}}$.
```

**Transfer**

The algorithm describes the sender transfers to recipient.

```
• inputs:
  – public parameters $pp$
  – two notes to spend $note_1$, $note_2$
  – sender's secret key $sk_A$
  – paths $path_i$ from commitment $cm_i$ to root $rt$, $i \in \{1, 2\}$
  – recipient's public key $pk_B$
  – the transfer value $v_B$
  – the Merkle root $rt$
• outputs: new notes $note_B$, $note_r$ and transfer transaction $\text{tx}_{\text{Transfer}}$
1) Parse $note_i = (pk_A, v_i, r_i), i \in \{1, 2\}$.
2) Compute $sn_i = \text{CRH}(sk_A, r_i), i \in \{1, 2\}$.
3) Sample two random numbers $r_3$, $r_4$.
4) Compute $cm_B = \text{Comm}(pk_B, v_B, r_3)$.
5) Set $note_B = (pk_B, v_B, r_3)$.
6) Compute $v_r = v_1 + v_2 - v_B$.
7) If $v_r \neq 0$, compute $cm_r = \text{Comm}(pk_A, v_r, r_4)$.
8) If $v_r \neq 0$, set $note_r = (pk_A, v_r, r_4)$.
9) Set $\vec{x} = (sn_1, sn_2, cm_B, cm_r, rt)$.
10) Set $\vec{\omega} = (note_1, note_2, note_B, note_r, sk_A, path_1, path_2)$.
11) Compute $\pi_{\text{Transfer}} = \text{Prove}(pk_{\text{Transfer}}, \vec{x}, \vec{w})$.
12) Set $\text{tx}_{\text{Transfer}} = (sn_1, sn_2, cm_B, cm_r, rt, \pi_{\text{Transfer}})$.
13) Output $note_B$, $note_r$ and $\text{tx}_{\text{Transfer}}$.
```

**Withdraw**

The algorithm describes how to withdraw from mixer.

```
• inputs:
  – public parameters $pp$
  – the note to redeem $note_B$
  – owner's secret key $sk_B$
  – path $path$ from commitment $cm_B$ to root $rt$
  – the account $repAcc$ to receive Ether
  – the Merkle root $rt$
• outputs: withdraw transaction $\text{tx}_{\text{Withdraw}}$
1) Parse $note_B = (pk_B, v_B, r_3)$.
2) Compute $sn = \text{CRH}(sk_B, r_3)$.
3) Set $\vec{x} = (sn, v_B, rt)$.
4) Set $\vec{\omega} = (pk_B, r_3, sk_B, path)$.
5) Compute $\pi_{\text{Withdraw}} = \text{Prove}(pk_{\text{Withdraw}}, \vec{x}, \vec{w})$.
6) Set $\text{tx}_{\text{Withdraw}} = (sn, v_B, rt, \pi_{\text{Withdraw}}, repAcc)$.
7) Output $\text{tx}_{\text{Withdraw}}$.
```

**VerifyTransaction**

The algorithm describes how mixer verifies transactions and makes corresponding operations.

```
• inputs:
  – public parameters $pp$
  – a transaction tx
• outputs: none
1) If given a deposit transaction $\text{tx} = \text{tx}_{\text{Deposit}}$:
    a) Parse $\text{tx}_{\text{Deposit}} = (cm, v, \pi_{\text{Deposit}})$.
    b) Verify the actual deposit value is v. (Revert if not).
    c) Set $\vec{x} = (cm, v)$.
    d) Compute $b = \text{Verify}(vk_{\text{Deposit}}, \vec{x}, \pi_{\text{Deposit}})$. (Revert if $b = 0$).
    e) Insert $cm$ into CMTree; and update CMTree.
2) If given a transfer transaction $\text{tx} = \text{tx}_{\text{Transfer}}$:
    a) Parse $\text{tx}_{\text{Transfer}} = (sn_1, sn_2, cm_B, cm_r, rt, \pi_{\text{Transfer}})$.
    b) Check $rt$ is in Roots. (Revert if not).
    c) Check neither $sn_1$ nor $sn_2$ is in SNSet. (Revert if not).
    d) Set $\vec{x} = (sn_1, sn_2, cm_B, cm_r, rt)$.
    e) Compute $b = \text{Verify}(vk_{\text{Transfer}}, \vec{x}, \pi_{\text{Transfer}})$. (Revert if $b = 0$).
    f) Insert $cm_B$ and $cm_r$ into CMTree; and update CMTree.
    g) Append $sn_1$ and $sn_2$ to SNSet.
3) If given a withdraw transaction $\text{tx} = \text{tx}_{\text{Withdraw}}$:
    a) Parse $\text{tx}_{\text{Withdraw}} = (sn, v_B, rt, \pi_{\text{Withdraw}}, repAcc)$.
    b) Check $rt$ is in Roots. (Revert if not).
    c) Check $sn$ is not in SNSet. (Revert if not).
    d) Set $\vec{x} = (sn, v_B, rt)$.
    e) Compute $b = \text{Verify}(vk_{\text{Withdraw}}, \vec{x}, \pi_{\text{Withdraw}})$. (Revert if $b = 0$).
    f) Transfer $v_B$ Ether to $repAcc$.
    g) Append $sn$ to SNSet.
```

Figure 1. Decentralized mixer mechanism.

### 4.1. Overview

For further realization of batch accounting, we apply the payment channel technology to the decentralized mixer. Specifically, some changes are made to the transfer operation of decentralized mixer. Here, we outline our construction in three incremental steps; the construction details see below. Note that the scheme of ZETH [11] is taken as base for the design of our proposed work.

**Attempt 1: the basic framework.** We first describe the basis framework of our scheme and point out its existing problems. Inspired by existing payment channel schemes in Bitcoin and Ethereum, where the channel corresponds to a multi-signature address or a smart contract respectively, we match the channel with a channel note denoted by $chnt$. In the OpenChannel phase, sender Alice utilizes unspent notes to create a channel note as a channel, i.e., $chnt_{AB} = (pk_A, v_{AB}, r_{AB})$, $pk_A$ denoting the owner of the channel, and $v_{AB}$ denoting the deposit locked in the channel later used to transfer to the recipient. In the OffchainTransfer phase, whenever Alice needs to transfer to Bob, she firstly creates a transaction $\mathrm{tx}^i_{\mathrm{OffchainTransfer}} = (chsn_{AB}, cm^i_B, cm^i_A, \pi^i_{\mathrm{OffchainTransfer}})$ and a note $note^i_B = (pk_B, v^i_B, r^i_B), i \in [1, n]$, the value $v^i_B$ representing the total amount Alice needs to transfer to Bob so far, and then transfers them to Bob through an anonymous network such as Tor [25]. The transactions between two parties are essentially the redistribution of the deposit in the channel. In the CloseChannel phase, either of them can post the latest transaction message to the blockchain network to get their money back.

However, the above draft may damage the interests of the recipient. The first problem with the attempt is that the channel note may be spent many times. For example, during the transaction between Alice and Bob, Alice utilizes the same channel note $chnt_{AB}$ to transact with Carl and generates $tx'_{\mathrm{OffchainTransfer}} = (chsn_{AB}, cm'_B, cm'_A, \pi'_{\mathrm{OffchainTransfer}})$ and a note $note_C = (pk_C, v_C, r_C)$. When Carl first closes the channel, the mixer verifies and adds $chsn_{AB}$ into SNSet. And then when Bob tries to close the channel, his transaction will be rejected because the channel note $chnt_{AB}$ has been spent. The second problem is when closing the channel, if the dishonest sender submits previous transactions not the latest transaction, i.e., $\mathrm{tx}^i_{\mathrm{OffchainTransfer}}, i < n$, not $\mathrm{tx}^n_{\mathrm{OffchainTransfer}}$, then the interest of recipient will be damaged due to the total transferred amount is included in the latest transaction.

**Attempt 2: maintaining the recipient's interests.** We make the second attempt to address the above challenges. To solve the double-spending problem, we require to define the channel's recipient $pk_B$, the channel note being $chnt_{AB} = (pk_A, v_{AB}, r_{AB}, pk_B)$. When $chnt_{AB}$ is spent to create new note $note_X$ (X is B or C), zero-knowledge proof $\pi_{\mathrm{OffchainTransfer}}$ is needed to prove the recipient in $chnt_{AB}$ is consistent with the owner of $note_X$. Therefore, the channel $chnt_{AB}$ is only used to transact with the recipient defined in the channel note, i.e., $pk_B$. To prevent the sender from broadcasting previous transactions, we rule that only the recipient can close the channel. The idea is accomplished by introducing difficult problems: (i) the recipient generates a difficult problem $diff_{AB}$ with a solution $x$, and only sends $diff_{AB}$ to the sender; (ii) the sender defines $diff_{AB}$ in the channel note, i.e., $chnt_{AB} = (pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB})$. It requires that only the one who knows the solution to the difficult problem can close the channel, so only the recipient can make it.

However, the second attempt may harm the interests of the sender. If the dishonest recipient never closes the channel, the balance $v_{AB} - v_B$ in the channel will never be returned to the sender. Not knowing the solution to the difficult problem, the sender has no choice but to wait the recipient to close the channel. This situation will harm the sender's interests.

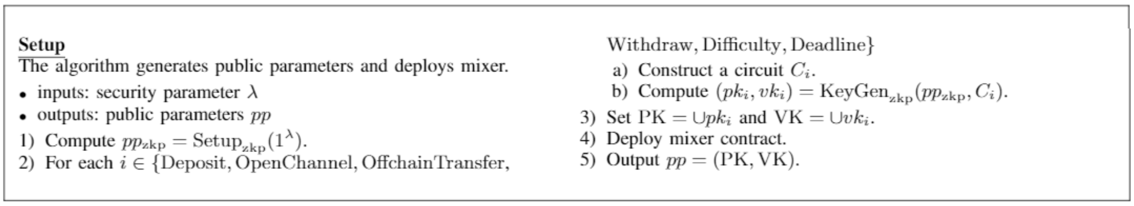
**Attempt 3: maintaining the sender's interests.** To overcome the above shortcoming, we set a deadline for a channel, which requires the recipient to close the channel before the deadline, otherwise the sender will have the right to close the channel. When Alice creates a channel, she defines a deadline $ddl_{AB}$ in the channel note, i.e., $chnt_{AB} = (pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB})$. When the deadline has passed, the sender can close the channel by proving that the current time is greater than the deadline. In order to prevent the sender from submitting the previous transaction when closing the channel, it is required that the recipient ought to close before the deadline.

In conclusion, we setup a one-way transaction channel from sender to recipient, which adds the defined recipient to prevent double-spending issues; applies difficult problems to ensure the interests of recipients; and uses the deadline to urge the recipient to close the channel on time.

## 4.2. Construction of DMC

In the following description, we detail the construction of DMC based on the mixing mechanism in Section 3. A DMC scheme Π is a tuple of algorithms (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw, VerifyTransaction, VerifyOffchainTransfer).

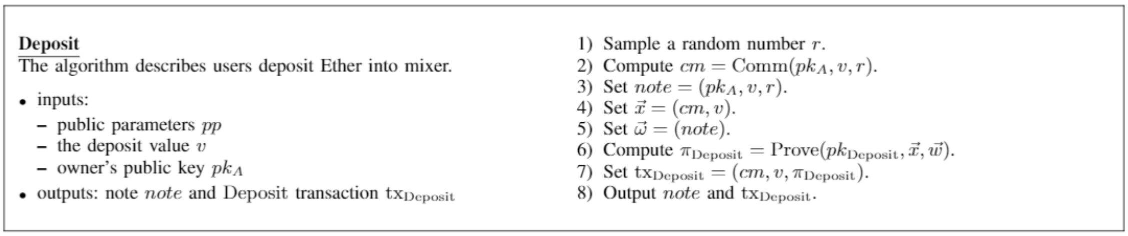
**Setup.** The algorithm generates a list of public parameters. To prove the validity of transactions, we build specific circuits which are taken to create keys for proof generation and verification. And the mixer contract is deployed on Ethereum. The detailed process proceeds as follows:



```
Setup
The algorithm generates public parameters and deploys mixer.
• inputs: security parameter $\lambda$
• outputs: public parameters $pp$
1) Compute $pp_{\mathrm{zkp}} = \mathrm{Setup}_{\mathrm{zkp}}(1^{\lambda})$.
2) For each $i \in \{$Deposit, OpenChannel, OffchainTransfer, Withdraw, Difficulty, Deadline$\}$
    a) Construct a circuit $C_i$.
    b) Compute $(pk_i, vk_i) = \mathrm{KeyGen}_{\mathrm{zkp}}(pp_{\mathrm{zkp}}, C_i)$.
3) Set $\mathrm{PK} = \cup pk_i$ and $\mathrm{VK} = \cup vk_i$.
4) Deploy mixer contract.
5) Output $pp = (\mathrm{PK}, \mathrm{VK})$.
```

**Deposit.** The algorithm builds a Deposit transaction $\mathrm{tx}_{\mathrm{Deposit}}$ to convert some Ether into an equivalent note. The transaction $\mathrm{tx}_{\mathrm{Deposit}}$ is composed of these variables.

• A new note commitment $cm$.
• A deposit value $v$.
• A zero-knowledge proof$\pi_{\mathrm{Deposit}}$, proving the following equation: $cm$ = $\mathrm{Comm}(pk_A, v, r)$.
The detailed process proceeds as follows:

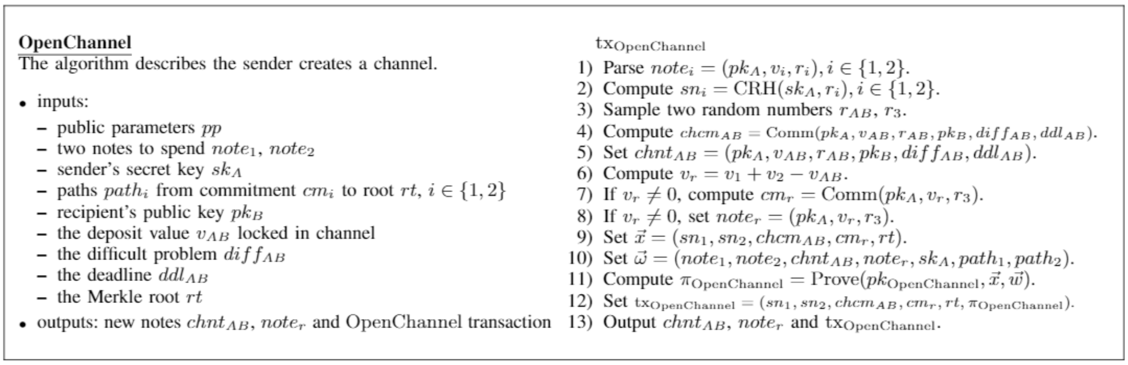


```
Deposit
The algorithm describes users deposit Ether into mixer.
• inputs:
    – public parameters $pp$
    – the deposit value $v$
    – owner's public key $pk_A$
• outputs: note $note$ and Deposit transaction $\mathrm{tx}_{\mathrm{Deposit}}$
1) Sample a random number $r$.
2) Compute $cm = \mathrm{Comm}(pk_A, v, r)$.
3) Set $note = (pk_A, v, r)$.
4) Set $\vec{x} = (cm, v)$.
5) Set $\vec{\omega} = (note)$.
6) Compute $\pi_{\mathrm{Deposit}} = \mathrm{Prove}(pk_{\mathrm{Deposit}}, \vec{x}, \vec{w})$.
7) Set $\mathrm{tx}_{\mathrm{Deposit}} = (cm, v, \pi_{\mathrm{Deposit}})$.
8) Output $note$ and $\mathrm{tx}_{\mathrm{Deposit}}$.
```

**OpenChannel.** The algorithm generates a OpenChannel transaction $\mathrm{tx}_{\mathrm{OpenChannel}}$, which utilizes $n$ (let $n$ =2) notes to create a channel note. The channel note is used as a channel for later transactions. The transaction $\mathrm{tx}_{\mathrm{OpenChannel}}$ is composed of these variables.

- Two serial numbers of spent notes $sn_1$ and $sn_2$.
- A channel note commitment $chcm_{AB}$.
- A balance commitment $cm_r$.
- The Merkle root $rt$.
- A zero-knowledge proof $\pi_{\mathsf{OpenChannel}}$, proving the following equations.
  - $cm_i$ = Comm($pk_A, v_i, r_i$), $i \in \{1,2\}$; $cm_r$ =Comm($pk_A, v_r, r_3$).
  - $chcm_{AB}$ = Comm($pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB}$).
  - $cm_i \in$ CMTree, $i \in \{1,2\}$.
  - $v_1 + v_2 = v_{AB} + v_3$.

The detailed process proceeds as follows:



**OpenChannel**
The algorithm describes the sender creates a channel.

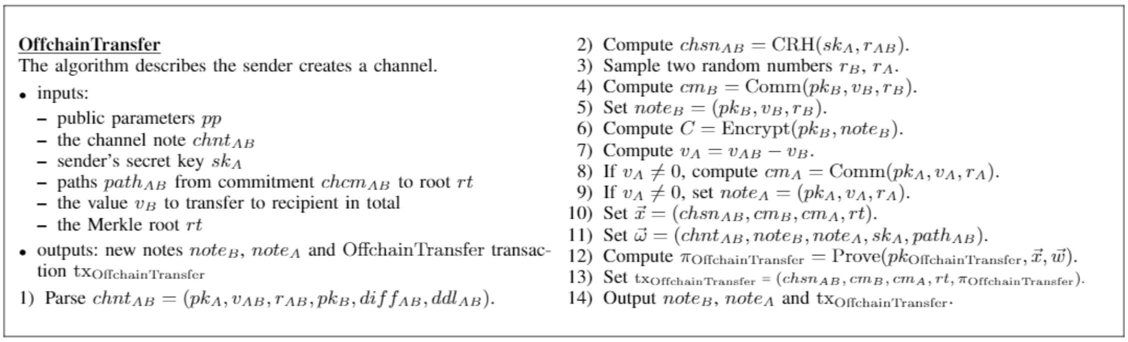
- inputs:
  - public parameters $pp$
  - two notes to spend $note_1$, $note_2$
  - sender's secret key $sk_A$
  - paths $path_i$ from commitment $cm_i$ to root $rt$, $i \in \{1, 2\}$
  - recipient's public key $pk_B$
  - the deposit value $v_{AB}$ locked in channel
  - the difficult problem $diff_{AB}$
  - the deadline $ddl_{AB}$
  - the Merkle root $rt$
- outputs: new notes $chnt_{AB}$, $note_r$ and OpenChannel transaction

```
tx_OpenChannel
1) Parse $note_i = (pk_A, v_i, r_i), i \in \{1, 2\}$.
2) Compute $sn_i = \mathrm{CRH}(sk_A, r_i), i \in \{1, 2\}$.
3) Sample two random numbers $r_{AB}$, $r_3$.
4) Compute $chcm_{AB} = \mathrm{Comm}(pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB})$.
5) Set $chnt_{AB} = (pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB})$.
6) Compute $v_r = v_1 + v_2 - v_{AB}$.
7) If $v_r \neq 0$, compute $cm_r = \mathrm{Comm}(pk_A, v_r, r_3)$.
8) If $v_r \neq 0$, set $note_r = (pk_A, v_r, r_3)$.
9) Set $\vec{x} = (sn_1, sn_2, chcm_{AB}, cm_r, rt)$.
10) Set $\vec{\omega} = (note_1, note_2, chnt_{AB}, note_r, sk_A, path_1, path_2)$.
11) Compute $\pi_{\mathrm{OpenChannel}} = \mathrm{Prove}(pk_{\mathrm{OpenChannel}}, \vec{x}, \vec{w})$.
12) Set $\mathrm{tx}_{\mathrm{OpenChannel}} = (sn_1, sn_2, chcm_{AB}, cm_r, rt, \pi_{\mathrm{OpenChannel}})$.
13) Output $chnt_{AB}$, $note_r$ and $\mathrm{tx}_{\mathrm{OpenChannel}}$.
```

**OffchainTransfer.** The algorithm utilizes the created channel to transfer to the recipient, generating an OffchainTransfer transaction $\mathrm{tx}_{\mathsf{OffchainTransfer}}$ and a new note $note_B$ which are sent to recipient through anonymous network. Every time the sender intends to transfer, she will execute the algorithm, redistributing the deposit in the channel. The transaction $\mathrm{tx}_{\mathsf{OffchainTransfer}}$ is composed of these variables.

- The serial number of channel note $chsn_{AB}$.
- A transfer commitment $cm_B$.
- A balance commitment $cm_A$.
- The Merkle root $rt$.
- A zero-knowledge proof$\pi_{\mathsf{OffchainTransfer}}$, proving the following equations.
  - $chcm_{AB}$ = Comm($pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB}$).
  - $cm_i$ = Comm($pk_i, v_i, r_i$), $i \in \{A, B\}$.
  - $sn_{AB}$ = CRH($sk_A, r_{AB}$).
  - $chcm_{AB} \in$ CMTree.
  - $v_{AB} = v_B + v_A$.

The detailed process proceeds as follows:



**OffchainTransfer**

The algorithm describes the sender creates a channel.

- inputs:
  - public parameters $pp$
  - the channel note $chnt_{AB}$
  - sender's secret key $sk_A$
  - paths $path_{AB}$ from commitment $chcm_{AB}$ to root $rt$
  - the value $v_B$ to transfer to recipient in total
  - the Merkle root $rt$
- outputs: new notes $note_B$, $note_A$ and OffchainTransfer transaction $\mathrm{tx}_{\mathrm{OffchainTransfer}}$

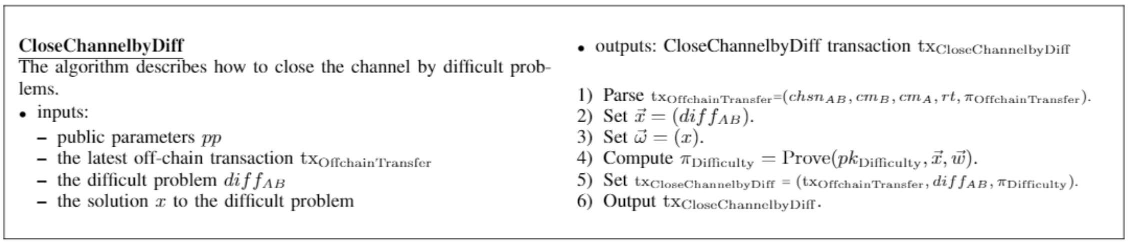
1) Parse $chnt_{AB} = (pk_A, v_{AB}, r_{AB}, pk_B, diff_{AB}, ddl_{AB})$.
2) Compute $chsn_{AB} = \mathrm{CRH}(sk_A, r_{AB})$.
3) Sample two random numbers $r_B$, $r_A$.
4) Compute $cm_B = \mathrm{Comm}(pk_B, v_B, r_B)$.
5) Set $note_B = (pk_B, v_B, r_B)$.
6) Compute $C = \mathrm{Encrypt}(pk_B, note_B)$.
7) Compute $v_A = v_{AB} - v_B$.
8) If $v_A \neq 0$, compute $cm_A = \mathrm{Comm}(pk_A, v_A, r_A)$.
9) If $v_A \neq 0$, set $note_A = (pk_A, v_A, r_A)$.
10) Set $\vec{x} = (chsn_{AB}, cm_B, cm_A, rt)$.
11) Set $\vec{\omega} = (chnt_{AB}, note_B, note_A, sk_A, path_{AB})$.
12) Compute $\pi_{\mathrm{OffchainTransfer}} = \mathrm{Prove}(pk_{\mathrm{OffchainTransfer}}, \vec{x}, \vec{w})$.
13) Set $\mathrm{tx}_{\mathrm{OffchainTransfer}} = (chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
14) Output $note_B$, $note_A$ and $\mathrm{tx}_{\mathrm{OffchainTransfer}}$.

**CloseChannel.** The CloseChannel operation is divided into two cases to discuss.

1) **CloseChannelbyDiff.** The algorithm describes the recipient generates a CloseChannel transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$. The recipient utilizes the solution to difficult problem to generate zero-knowledge proof, and posts $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$ to blockchain network before the deadline of the channel. The transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$ is composed of the next variables.

- The latest off-chain transaction $\mathrm{tx}_{\mathrm{OffchainTransfer}}$.
- The difficult problem $diff_{AB}$.
- A zero-knowledge proof $\pi_{\mathrm{Difficulty}}$, proving the following equation: $x$ is a solution to $diff_{AB}$.

The detailed process proceeds as follows:



**CloseChannelbyDiff**

The algorithm describes how to close the channel by difficult problems.

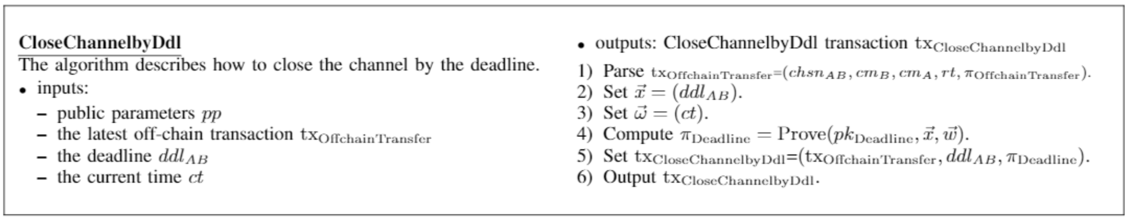
- inputs:
  - public parameters $pp$
  - the latest off-chain transaction $\mathrm{tx}_{\mathrm{OffchainTransfer}}$
  - the difficult problem $diff_{AB}$
  - the solution $x$ to the difficult problem
- outputs: CloseChannelbyDiff transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$

1) Parse $\mathrm{tx}_{\mathrm{OffchainTransfer}}{=}(chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
2) Set $\vec{x} = (diff_{AB})$.
3) Set $\vec{\omega} = (x)$.
4) Compute $\pi_{\mathrm{Difficulty}} = \mathrm{Prove}(pk_{\mathrm{Difficulty}}, \vec{x}, \vec{w})$.
5) Set $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}} = (\mathrm{tx}_{\mathrm{OffchainTransfer}}, diff_{AB}, \pi_{\mathrm{Difficulty}})$.
6) Output $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$.

2) **CloseChannelbyDdl.** The algorithm describes the sender generates a CloseChannel transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$. If the recipient does not close the channel in time, the sender can do by proving to mixer contract that the deadline has passed. The transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$ is composed of the next variables.

- The latest off-chain transaction$\mathrm{tx}_{\mathrm{OffchainTransfer}}$.
- The difficult problem $ddl_{AB}$.
- A zero-knowledge proof $\pi_{\mathrm{Deadline}}$, proving the following equation: $ct > ddl_{AB}$.

The detailed process proceeds as follows:



**CloseChannelbyDdl**

The algorithm describes how to close the channel by the deadline.

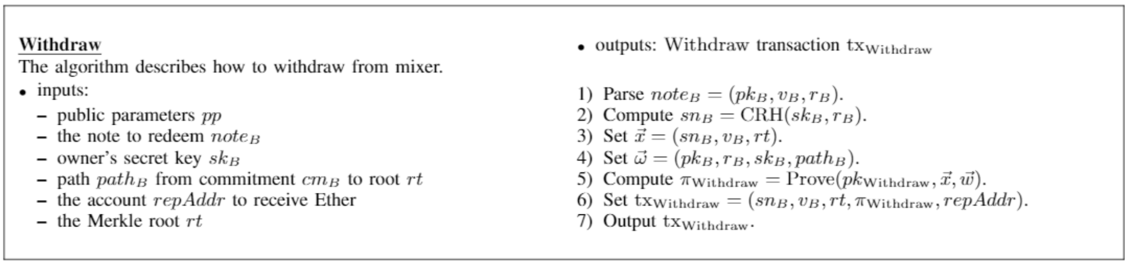
- inputs:
  - public parameters $pp$
  - the latest off-chain transaction $\mathrm{tx}_{\mathrm{OffchainTransfer}}$
  - the deadline $ddl_{AB}$
  - the current time $ct$
- outputs: CloseChannelbyDdl transaction $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$

1) Parse $\mathrm{tx}_{\mathrm{OffchainTransfer}}{=}(chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
2) Set $\vec{x} = (ddl_{AB})$.
3) Set $\vec{\omega} = (ct)$.
4) Compute $\pi_{\mathrm{Deadline}} = \mathrm{Prove}(pk_{\mathrm{Deadline}}, \vec{x}, \vec{w})$.
5) Set $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}{=}(\mathrm{tx}_{\mathrm{OffchainTransfer}}, ddl_{AB}, \pi_{\mathrm{Deadline}})$.
6) Output $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$.

**Withdraw.** The algorithm constructs a Withdraw transaction to redeem a note into equivalent Ether. The transaction $\mathrm{tx}_{\mathrm{Withdraw}}$ is composed of the following variables.

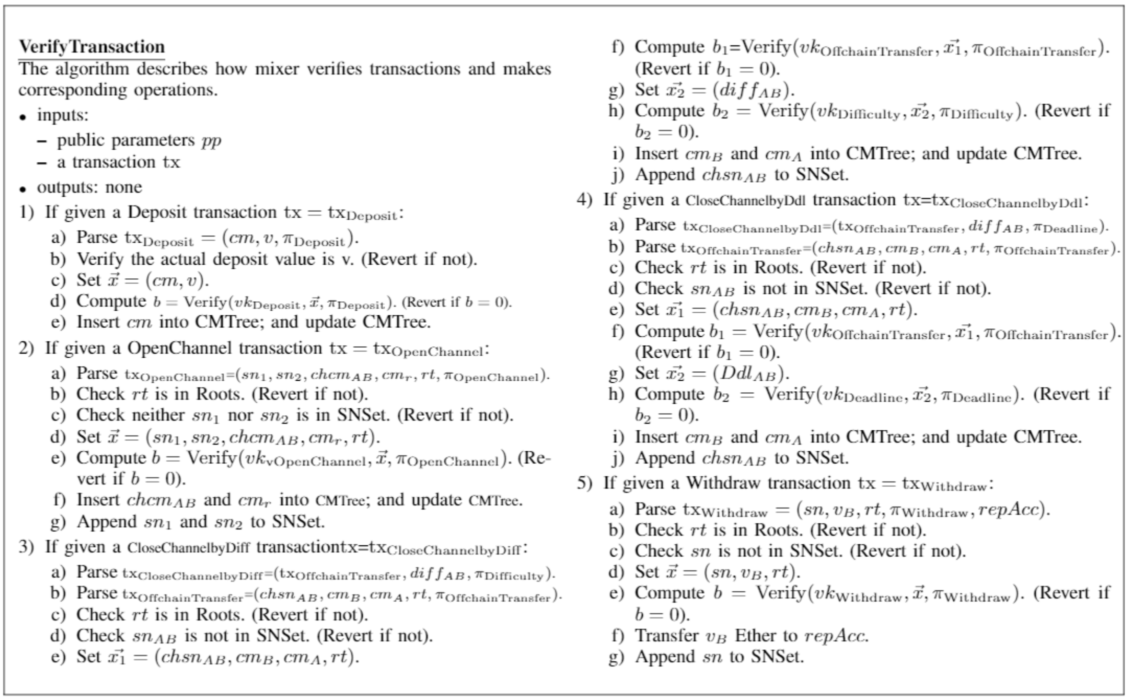
- The serial number$sn_B$.
- The withdraw value$v_B$.
- The Merkle root $rt$.
- An account$repAddr$ to receive Ether.
- A zero-knowledge proof$\pi_{\mathrm{Withdraw}}$, proving the following equations.
  - $cm_B$ = Comm($pk_B, v_B, r_B$).
  - $sn_B$ = CRH($sk_B, r_B$).
  - $cm_B \in$ CMTree.

The detailed process proceeds as follows:



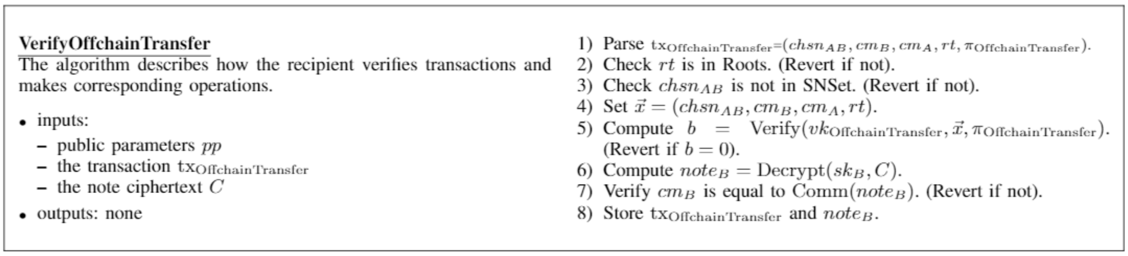
```
Withdraw
The algorithm describes how to withdraw from mixer.
• inputs:
  – public parameters $pp$
  – the note to redeem $note_B$
  – owner's secret key $sk_B$
  – path $path_B$ from commitment $cm_B$ to root $rt$
  – the account $repAddr$ to receive Ether
  – the Merkle root $rt$
• outputs: Withdraw transaction $\mathrm{tx}_{\mathrm{Withdraw}}$

1) Parse $note_B = (pk_B, v_B, r_B)$.
2) Compute $sn_B = \mathrm{CRH}(sk_B, r_B)$.
3) Set $\vec{x} = (sn_B, v_B, rt)$.
4) Set $\vec{\omega} = (pk_B, r_B, sk_B, path_B)$.
5) Compute $\pi_{\mathrm{Withdraw}} = \mathrm{Prove}(pk_{\mathrm{Withdraw}}, \vec{x}, \vec{w})$.
6) Set $\mathrm{tx}_{\mathrm{Withdraw}} = (sn_B, v_B, rt, \pi_{\mathrm{Withdraw}}, repAddr)$.
7) Output $\mathrm{tx}_{\mathrm{Withdraw}}$.
```

**VerifyTransaction.** This algorithm checks by the mixer contract all transactions except OffchainTransfer transactions. The contract verifies the uniqueness of serial numbers, the correctness of note commitments and the validity of Merkle root. If all checks are satisfied, it will perform corresponding operations: (i) add commitments into CMTree; (ii) append serial numbers to SNSet; or (iii) transfer Ether to defined account. The detailed process proceeds as follows:



```
VerifyTransaction
The algorithm describes how mixer verifies transactions and makes corresponding operations.
• inputs:
  – public parameters $pp$
  – a transaction tx
• outputs: none
1) If given a Deposit transaction $\mathrm{tx} = \mathrm{tx}_{\mathrm{Deposit}}$:
    a) Parse $\mathrm{tx}_{\mathrm{Deposit}} = (cm, v, \pi_{\mathrm{Deposit}})$.
    b) Verify the actual deposit value is v. (Revert if not).
    c) Set $\vec{x} = (cm, v)$.
    d) Compute $b = \mathrm{Verify}(vk_{\mathrm{Deposit}}, \vec{x}, \pi_{\mathrm{Deposit}})$. (Revert if $b = 0$).
    e) Insert $cm$ into CMTree; and update CMTree.
2) If given a OpenChannel transaction $\mathrm{tx} = \mathrm{tx}_{\mathrm{OpenChannel}}$:
    a) Parse $\mathrm{tx}_{\mathrm{OpenChannel}}$=$(sn_1, sn_2, chcm_{AB}, cm_r, rt, \pi_{\mathrm{OpenChannel}})$.
    b) Check $rt$ is in Roots. (Revert if not).
    c) Check neither $sn_1$ nor $sn_2$ is in SNSet. (Revert if not).
    d) Set $\vec{x} = (sn_1, sn_2, chcm_{AB}, cm_r, rt)$.
    e) Compute $b = \mathrm{Verify}(vk_{\mathrm{vOpenChannel}}, \vec{x}, \pi_{\mathrm{OpenChannel}})$. (Revert if $b = 0$).
    f) Insert $chcm_{AB}$ and $cm_r$ into CMTree; and update CMTree.
    g) Append $sn_1$ and $sn_2$ to SNSet.
3) If given a CloseChannelbyDiff transactiontx=$\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$:
    a) Parse $\mathrm{tx}_{\mathrm{CloseChannelbyDiff}}$=$(\mathrm{tx}_{\mathrm{OffchainTransfer}}, diff_{AB}, \pi_{\mathrm{Difficulty}})$.
    b) Parse $\mathrm{tx}_{\mathrm{OffchainTransfer}}$=$(chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
    c) Check $rt$ is in Roots. (Revert if not).
    d) Check $sn_{AB}$ is not in SNSet. (Revert if not).
    e) Set $\vec{x_1} = (chsn_{AB}, cm_B, cm_A, rt)$.
    f) Compute $b_1$=$\mathrm{Verify}(vk_{\mathrm{OffchainTransfer}}, \vec{x_1}, \pi_{\mathrm{OffchainTransfer}})$. (Revert if $b_1 = 0$).
    g) Set $\vec{x_2} = (diff_{AB})$.
    h) Compute $b_2 = \mathrm{Verify}(vk_{\mathrm{Difficulty}}, \vec{x_2}, \pi_{\mathrm{Difficulty}})$. (Revert if $b_2 = 0$).
    i) Insert $cm_B$ and $cm_A$ into CMTree; and update CMTree.
    j) Append $chsn_{AB}$ to SNSet.
4) If given a CloseChannelbyDdl transaction tx=$\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$:
    a) Parse $\mathrm{tx}_{\mathrm{CloseChannelbyDdl}}$=$(\mathrm{tx}_{\mathrm{OffchainTransfer}}, diff_{AB}, \pi_{\mathrm{Deadline}})$.
    b) Parse $\mathrm{tx}_{\mathrm{OffchainTransfer}}$=$(chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
    c) Check $rt$ is in Roots. (Revert if not).
    d) Check $sn_{AB}$ is not in SNSet. (Revert if not).
    e) Set $\vec{x_1} = (chsn_{AB}, cm_B, cm_A, rt)$.
    f) Compute $b_1 = \mathrm{Verify}(vk_{\mathrm{OffchainTransfer}}, \vec{x_1}, \pi_{\mathrm{OffchainTransfer}})$. (Revert if $b_1 = 0$).
    g) Set $\vec{x_2} = (Ddl_{AB})$.
    h) Compute $b_2 = \mathrm{Verify}(vk_{\mathrm{Deadline}}, \vec{x_2}, \pi_{\mathrm{Deadline}})$. (Revert if $b_2 = 0$).
    i) Insert $cm_B$ and $cm_A$ into CMTree; and update CMTree.
    j) Append $chsn_{AB}$ to SNSet.
5) If given a Withdraw transaction $\mathrm{tx} = \mathrm{tx}_{\mathrm{Withdraw}}$:
    a) Parse $\mathrm{tx}_{\mathrm{Withdraw}} = (sn, v_B, rt, \pi_{\mathrm{Withdraw}}, repAcc)$.
    b) Check $rt$ is in Roots. (Revert if not).
    c) Check $sn$ is not in SNSet. (Revert if not).
    d) Set $\vec{x} = (sn, v_B, rt)$.
    e) Compute $b = \mathrm{Verify}(vk_{\mathrm{Withdraw}}, \vec{x}, \pi_{\mathrm{Withdraw}})$. (Revert if $b = 0$).
    f) Transfer $v_B$ Ether to $repAcc$.
    g) Append $sn$ to SNSet.
```

**VerifyOffchainTransfer.** This algorithm checks OffchainTransfer transactions by the recipient. If passed, the transaction and note messages are stored. The detailed process proceeds as follows:



**VerifyOffchainTransfer**

The algorithm describes how the recipient verifies transactions and makes corresponding operations.

- inputs:
  - public parameters $pp$
  - the transaction $\mathrm{tx}_{\mathrm{OffchainTransfer}}$
  - the note ciphertext $C$
- outputs: none

```
1) Parse $\mathrm{tx}_{\mathrm{OffchainTransfer}}=(chsn_{AB}, cm_B, cm_A, rt, \pi_{\mathrm{OffchainTransfer}})$.
2) Check $rt$ is in Roots. (Revert if not).
3) Check $chsn_{AB}$ is not in SNSet. (Revert if not).
4) Set $\vec{x} = (chsn_{AB}, cm_B, cm_A, rt)$.
5) Compute $b = \mathrm{Verify}(vk_{\mathrm{OffchainTransfer}}, \vec{x}, \pi_{\mathrm{OffchainTransfer}})$.
   (Revert if $b = 0$).
6) Compute $note_B = \mathrm{Decrypt}(sk_B, C)$.
7) Verify $cm_B$ is equal to $\mathrm{Comm}(note_B)$. (Revert if not).
8) Store $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ and $note_B$.
```

## 4.3. Security of DMC

Following the security model defined in the Zerocash [17] and BlockMaze [7], we define two secure properties of DMC: transaction unlinkability and overdraft safety. The formal security definitions are provided in Appendix A.

**Definition 1** (Security)**.** A DMC scheme is secure if it satisfies transaction unlinkability and overdraft safety as defined in the experiments in Figure 2. (Note, in $\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{TR-UL}}(\lambda)$, participant Of denotes sender or recipient of transactions, addrOf denotes addresses of the adversary. In $\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{OD-SF}}(\lambda)$, InOut is used to compute the income and outcome related to the account of $\mathcal{A}$.)

1) **Transaction unlinkability.** The property, defined by the TR-UL experiment, means that no probabilistic polynomial-time (PPT) adversary can recognize the transaction linkage between the sender and recipient. The scheme Π is transaction unlinkable if

$$\Pr[\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{TR-UL}}(\lambda) = 1] \leq negl(\lambda) \tag{1}$$

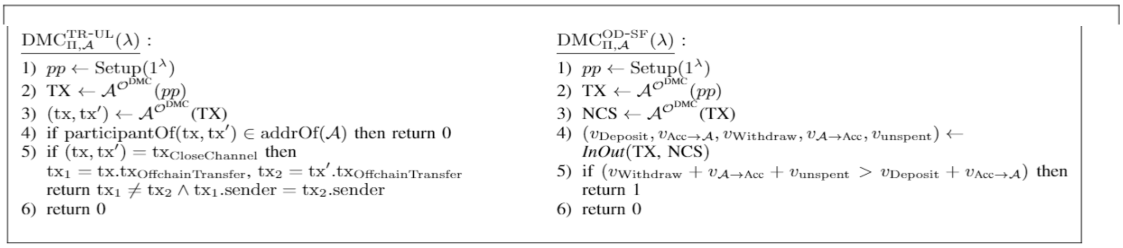
where $\Pr[\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{TR-UL}}(\lambda) = 1]$ represents the winning probability of $\mathcal{A}$ in the TR-UL experiment.

2) **Overdraft Safety.** The property, formalized in the OD-SF experiment, shows that no PPT adversary can spend more coins than what he deposits and receives from others. The scheme Π is overdraft safe if

$$\Pr[\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{OD-SF}}(\lambda) = 1] \leq negl(\lambda) \tag{2}$$

where $\Pr[\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{OD-SF}}(\lambda) = 1]$ means the winning probability of $\mathcal{A}$ in the OD-SF experiment.

**Theorem 1.** The tuple Π = (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw, VerifyTransaction, VerifyOffchainTransfer) is a secure DMC scheme. (The proof is provided in Appendix B.)



```
$\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{TR\text{-}UL}}(\lambda)$ :
1) $pp \leftarrow \mathrm{Setup}(1^\lambda)$
2) $\mathrm{TX} \leftarrow \mathcal{A}^{\mathcal{O}^{\mathrm{DMC}}}(pp)$
3) $(\mathrm{tx}, \mathrm{tx}') \leftarrow \mathcal{A}^{\mathcal{O}^{\mathrm{DMC}}}(\mathrm{TX})$
4) if $\mathrm{participantOf}(\mathrm{tx}, \mathrm{tx}') \in \mathrm{addrOf}(\mathcal{A})$ then return 0
5) if $(\mathrm{tx}, \mathrm{tx}') = \mathrm{tx}_{\mathrm{CloseChannel}}$ then
   $\mathrm{tx}_1 = \mathrm{tx.tx}_{\mathrm{OffchainTransfer}}$, $\mathrm{tx}_2 = \mathrm{tx}'.\mathrm{tx}_{\mathrm{OffchainTransfer}}$
   return $\mathrm{tx}_1 \neq \mathrm{tx}_2 \wedge \mathrm{tx}_1.\mathrm{sender} = \mathrm{tx}_2.\mathrm{sender}$
6) return 0
```

```
$\mathrm{DMC}_{\Pi,\mathcal{A}}^{\mathrm{OD\text{-}SF}}(\lambda)$ :
1) $pp \leftarrow \mathrm{Setup}(1^\lambda)$
2) $\mathrm{TX} \leftarrow \mathcal{A}^{\mathcal{O}^{\mathrm{DMC}}}(pp)$
3) $\mathrm{NCS} \leftarrow \mathcal{A}^{\mathcal{O}^{\mathrm{DMC}}}(\mathrm{TX})$
4) $(v_{\mathrm{Deposit}}, v_{\mathrm{Acc}\rightarrow\mathcal{A}}, v_{\mathrm{Withdraw}}, v_{\mathcal{A}\rightarrow\mathrm{Acc}}, v_{\mathrm{unspent}}) \leftarrow$
   $\mathit{InOut}(\mathrm{TX}, \mathrm{NCS})$
5) if $(v_{\mathrm{Withdraw}} + v_{\mathcal{A}\rightarrow\mathrm{Acc}} + v_{\mathrm{unspent}} > v_{\mathrm{Deposit}} + v_{\mathrm{Acc}\rightarrow\mathcal{A}})$ then
   return 1
6) return 0
```

Figure 2. The transaction unlinkability and overdraft safety experiment for DMC.

## 5. IMPLEMENT AND PERFORMANCE EVALUATION

In this section, we first instantiate cryptographic building blocks, and then implement our DMC scheme as a specific mixer contract. At last, we conduct comprehensive experiments to evaluate its performance. Our source code will be available at https://github.com/LS291730/DMC.

### 5.1. Cryptographic Building Blocks

As for collision-resistant cryptographic hash function (CRH), we choose MiMC [26] hash. Compared to other hash functions (e.g. SHA-256 and Keccak), MiMC is friendly to arithmetic circuits, creating lower number of constraints and operations. The commitment scheme is directly instantiated using MiMC hash function, same as the difficult problem and the hash algorithm used in the Merkle tree.

We use the *eciespy*, Elliptic Curve Integrated Encryption Scheme (ECIES) for secp256k1 in Python, for the encryption scheme. In this scheme, transaction messages are directly encrypted with Ethereum public key and decrypted with Ethereum private key.

We take Groth16 [27] as our instance of zk-SNARKs due to its efficiency in term of proof size and verification time. Groth16 is an excellent zk-SNARK proving scheme which, compared with other schemes, has a smaller proof size with fixed 256 bytes and a faster verification speed at millisecond level. Note that in our implementation, the setup phase of zero-knowledge proof is created by a trusted third party.

### 5.2. Implementation

We implement our DMC scheme based on zk-SNARK tools (e.g. *circom* [27], a low-level circuit language and a compiler, and *snarkjs* [28], a JavaScript implementation of zk-SNARKs), and Ethereum tools (e.g. Web3.py [29], a Python library for interacting with Ethereum, and Ganache, a local Ethereum blockchain which generates some virtual accounts that we can use during development.). For user algorithms, written by Python, they allow users to create transactions. Users can send via Web3.py these transactions to the blockchain network, interacting with mixer contract. In addition, we use *cir com* to construct arithmetic circuits; and later apply zero-knowledge tool *snarkjs* to generate and validate zero-knowledge proofs. For mixer contract, it is programmed by Solidity, compiled to EVM bytecode and later deployed on Ganache. The functions in mixer, such as *deposit*, *openChannel*, *closeChannel* and *withdraw*, will verify corresponding transactions and make corresponding operations. Note that DMC currently only supports private transfer of Ether, but can later be expanded to support various tokens, such as ERC-20 and ERC-721 tokens.

**zk-SNARKs for DMC transactions.** For these transactions in DMC (i.e., Deposit, OpenChannel, OffchainTransfer, CloseChannel and Withdraw), we utilize zk-SNARKs to construct zero-knowledge proofs according to their respective circuits. The common reference string (CRS) related to each zero-knowledge proof is generated by a trusted third party and later destroyed to guarantee security. And the generated key pairs for proof generation and verification are public, available to users and mixer contract.

### 5.3. Performance Evaluation

We conduct experiments to evaluate the performance of the proposed mixing scheme. First of all, we estimate the performance of zero-knowledge proofs. Then, we measure the gas cost consumed by transactions involved in our scheme and analyse the main factors for the gas cost. At last, we analyse the decrease in transaction cost and latency. Note that the following experiments are executed 10 times and we take the average value.

We now consider the performance of zero-knowledge proofs in terms of setup time, key pair size and proof generation/verification time. These are summarized in table 1. Note that the generator time refers to the time executing both $\mathsf{Setup}_{\mathsf{zkp}}$ and $\mathsf{KeyGen}_{\mathsf{zkp}}$ algorithms for each of zero-knowledge proofs. For each proof, the generator time depends on the complexity of circuit (e.g., circuit $\mathcal{C}_{\mathsf{Deposit}}$ contains 1 MiMC gadget while circuit $\mathcal{C}_{\mathsf{OpenChannel}}$ contains 6 MiMC gasgets and 2 Merkle tree gadgets), the same as the generation time. Furthermore, the generator time is linearly dependent on the size of the proving key. Instead, the size of verification key and the time of proof verification are maintained stable, irrelevant to the circuit's complexity.

Table 1. The performance of zero-knowledge proofs.

| **ZKP** | **Generator time** | **Proving key size** | **Verification key size** | **Proof generation time** | **Proof verification time** |
|---|---|---|---|---|---|
| Deposit | 49.4s | 1.0MB | 640B | 0.77s | 0.376s |
| OpenChannel | 6m51s | 6.2MB | 832B | 1.53s | 0.368s |
| OffchainTransfer | 3m50s | 4.1MB | 768B | 1.33s | 0.371s |
| Withdraw | 52.3s | 1.8MB | 704B | 0.91s | 0.360s |

The cost to deploy the DMC mixing contract is 2,294,567 gas. Table 2 shows the gas cost consumed by these transactions sent to the mixer contract. For Deposit transactions, it consumes the first sender 1,090,661 gas to process the first transaction, but 610,653 (given in the table) for the other transactions. The extra gas is costed to set storage in the EVM. A majority of the gas cost lies in two chief operations: the verification of zero-knowledge proofs and the update of the Merkle tree CMTree, both of which cost approximately 200,000 gas. The numbers of verification and update operations involved in each transaction are given in Table 3. As seen from the table, the CloseChannel transaction costs the most gas while the Withdraw transaction costs the least because the former has two verification and update operations respectively while the latter needs to verify zero-knowledge proof only once.

Table 2. The gas cost of transactions for interacting with mixer contract.

| **Transaction** | **Gas cost** | **#VerifyProof** | **#UpdateTree** |
|---|---|---|---|
| $\mathsf{tx}_{\mathsf{Deposit}}$ | 610,653 | 1 | 1 |
| $\mathsf{tx}_{\mathsf{OpenChannel}}$ | 919,108 | 1 | 2 |
| $\mathsf{tx}_{\mathsf{CloseChannelbyDiff}}$ | 1,118,862 | 2 | 2 |
| $\mathsf{tx}_{\mathsf{CloseChannelbyDdl}}$ | 1,126,453 | 2 | 2 |
| $\mathsf{tx}_{\mathsf{Withdraw}}$ | 288,421 | 1 | 0 |

Compared with related work, the gas cost of related decentralized mixing schemes is given in Table 3 (Note that $k$ and $j$ denote the number of participants in the ring signature or shuffle and the number of malicious shuffles respectively). In the first two schemes, one deposit transaction corresponds to one withdrawal transaction without an extra transfer operation. When withdrawing from mixer contract, Möbius and Miximus utilize verifiable ring signature and zero-knowledge proof respectively to create the withdrawal transaction. So, the cost of withdrawal operation in Möbius grows linearly with the number of participants in ring signature, and that in Miximus is relatively high for proof verification. In MixEth, before withdrawing, several shuffle operations are required to perform to break the transaction relationship, and then recipients need to check the correctness of the preceding shuffles. The latter two schemes use existing deposits in the mixer to transfer to recipients. Zether costs more gas for applying encryption scheme and Σ-Bullets to transfer. While in our scheme DMC, because the sender sends transaction messages directly to the recipient through anonymous network, there is no need to interact with Ethereum network except one transaction to open a channel (919,108 gas) and another to close the channel (1,118,862 gas).

Table 3. Comparison between gas costs of different decentralized mixing schemes.

| **Mixer** | **Deposit** | **Withdraw** | **Transfer** | **Total** |
|---|---|---|---|---|
| Möbius [9] | 76,123 | 335,714$k$ | - | 1,418,979 |
| Miximus [10] | 732,815 | 1,903,305 | - | 2,636,120 |
| MixEth [12] | 99,254 | 113,265 | 366,216+10,000$k$+227,563$j$ | 1,528,987 |
| Zether [8] | 260,000 | 384,000 | 7,188,000 | 7,832,000 |
| DMC | 610,653 | 288,421 | 0 | 3,547,697/$n$ |

For a complete transaction between two parties, the total transaction cost includes depositing, transferring (if exists) and withdrawing operations. Here, we set both the number of participants in ring signature/shuffle and the number of malicious shuffles to 4, i.e., $k$ = 4 and $j$ = 4. In our scheme, we suppose that the sender and recipient make $n$ off-chain transactions in total via the channel, i.e., the number of transactions in batch is $n$. Since there is no need for consensus for these transactions, the total cost of $n$ transactions only covers depositing, opening and closing of channels and withdrawing operations. By comparison with other schemes, the average cost of a transaction is 3,547,697/$n$, which is approximately 1/$n$ of others. For the same reason, the transaction latency is also about 1/$n$ of other schemes. Because these $n$ transactions are free from the effect of the underlying block generation mechanism and network congestion. On the other hand, the communication delay of anonymous network, compared with Ethereum network, is negligible.

Overall, the experimental results show that our proposed scheme is feasible on Ethereum. From theoretical and experimental analysis of DMC, we obtain that the average transaction cost and transaction latency are both about 1/$n$ of other mixing schemes.

## 6. RELATED WORK

Currently, transaction privacy-preserving schemes mainly include coin mixer, ring signature, zero-knowledge proof and trusted computation, etc. However, in this paper, we only focus the mixing schemes, more specifically, the decentralized ones. First, based on the balance model, the decentralized mixing schemes can be divided into account-based model [8] and UTXO-based model [11]. On the other hand, these schemes can also be classified to any mixing amount [8], [11] and fixed mixing amount [9], [12]. Some decentralized mixing schemes in the literature are briefly introduced as follows.

Mobius [9] presents a decentralized mixer, which only supports for transactions of fixed denominations. The scheme just involves deposit and withdrawal operations. To deposit, the sender derives a new stealth address to hide the recipient. When withdrawing, the recipient generates verifiable ring signature to prove his ownership of coins. The ring signature obscures recipients, however the gas cost consumed by signature verification increases linearly with the size of recipient set.

Zether [8] proposes an account-based coin mixer, i.e., users' deposits are placed in accounts in the form of ciphertext. When transferring, it utilizes homomorphic encryption to hide transaction amount, uses an anonymous account set to hide the sender and recipient, and exploits zero-knowledge proof to ensure the validity of transactions. Though there is no need for trusted setup, the overhead scales linearly with the size of the anonymous set.

MixEth [12], a trustless coin mixer, does not rely on a trusted setup. It uses shuffling method to break the relationship of two parties coin mixing, achieving strong notions of anonymity. Shuffling and challenging rounds are made in turns. Computing the shuffle is done off-chain, verifying the correctness of the new shuffling on-chain.

## 7. CONCLUSIONS

The decentralized mixing scheme suffers from high transaction cost for complex operations and expensive computing resources and long transaction latency for block generation mechanism and network congestion. In this paper, we adopt the idea of batch accounting to improve efficiency, reducing the transaction cost and latency issues. As the technical support behind batch accounting, we introduce payment channel technology into the mixing scheme and propose a decentralized mixer with channel called DMC. DMC works well in combination with the advantages of decentralized mixer and payment channel, decreasing the transaction cost and latency, while breaking the transaction relationship and hiding the transaction value. By the created channel, the transactions between two parties are transmitted through anonymous network. Since these off-chain transactions avoid network consensus, we achieve the decrease in transaction cost and latency.

Future scope of our proposed scheme is (i) There is need of utilizing secure multi-party computation (MPC) to avoid the trusted setup of zero-knowledge proof. (ii) The channel between the two sides needs to be expanded from one-way to two-way. (iii) The mixer with channel scheme will be likely to scale into other blockchain system, such as Bitcoin. (iv) The privacy protection method, not just the mixing scheme, can be combined with the two-layer scaling solutions to improve performance.

## ACKNOWLEDGEMENTS

This work is partly supported by the National Key Research and Development Program of China (No. 2020YFB1005900), the Research Fund of Guangxi Key Laboratory of Trusted Software (No. kx201906), and the Research Fund of Guangxi Key Lab of Multi-source Information Mining \& Security (No. MIMS20-07).

## REFERENCES

[1] Béres, F., Seres, I. A., Benczúr, A. A., &Quintyne-Collins, M. (2020). “Blockchain is watching you: Profiling and deanonymizing ethereum users”. *arXiv preprint arXiv*:2005.14051.

[2] Klusman, R., &Dijkhuizen, T. (2018). “Deanonymisation in ethereum using existing methods for bitcoin”.

[3] Victor, F. (2020). “Address clustering heuristics for Ethereum”. *In International Conference on Financial Cryptography and Data Security*, pp. 617-633.

[4] Payette, J., Schwager, S., & Murphy, J. (2017). “Characterizing the ethereum address space”.

[5] Chan, W., & Olmsted, A. (2017). “Ethereum transaction graph analysis”. *In 2017 12th international conference for internet technology and secured transactions*, pp. 498-500.

[6] Ma, S., Deng, Y., He, D., Zhang, J., &Xie, X. (2020). “An efficient NIZK scheme for privacy-preserving transactions over account-model blockchain”. *IEEE Transactions on Dependable and Secure Computing*, Vol. 18, No. 2,pp. 641-651.

[7] Guan, Z., Wan, Z., Yang, Y., Zhou, Y., & Huang, B. (2020). "Blockmaze: An efficient privacy-preserving account-model blockchain based on zk-SNARKs". *IEEE Transactions on Dependable and Secure Computing.*

[8] Bünz, B., Agrawal, S., Zamani, M., &Boneh, D. (2020). "Zether: Towards privacy in a smart contract world". *In International Conference on Financial Cryptography and Data Security*, pp. 423-443.

[9] Meiklejohn, S., & Mercer, R. (2018). "Möbius: Trustless Tumbling for Transaction Privacy. Proceedings on Privacy Enhancing Technologies", pp. 105-121.

[10] Barry Whitehat. (2018). "Miximus: zksnark-based trustless mixing for Ethereum". https://github.com/barryWhiteHat/miximus.

[11] Rondelet, A., & Zajac, M. (2019). "Zeth: On integrating zerocash on ethereum". *arXiv preprint arXiv*:1904.00905.

[12] Seres, I. A., Nagy, D. A., Buckland, C., &Burcsi, P. (2019). "Mixeth: efficient, trustless coin mixing service for ethereum". *In International Conference on Blockchain Economics, Security and Protocols.*

[13] Zachary J. Williamson. (2018). The AZTEC protocol. Available at: https://github.com/Aztec Protocol/AZTECblob/master/AZTEC.pdf.

[14] Nightfall implementation. Available at: https://github.com/EYBlockchain/nightfall.

[15] Tornado cash implementation. Available at: https://github.com/tornadocash/tornado-core.

[16] Gervais, A., Karame, G. O., Wüst, K., Glykantzis, V., Ritzdorf, H., &Capkun, S. (2016). "On the security and performance of proof of work blockchains". *In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. pp. 3-16.

[17] Sasson, E. B., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., &Virza, M. (2014). "Zerocash: Decentralized anonymous payments from bitcoin". *In 2014 IEEE Symposium on Security and Privacy*. pp. 459-474.

[18] Gavin Wood. (2014). "Ethereum: A Secure Decentralised Generalised Transaction Ledger". *Ethereum project yellow paper*.

[19] Tremback, J., & Hess, Z. (2015). "Universal payment channels".

[20] Dziembowski, S., Faust, S., &Hostáková, K. (2018). "General state channel networks". *In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. pp. 949-966.

[21] Poon, J., &Dryja, T. (2016). "The bitcoin lightning network: Scalable off-chain instant payments". Available at: https:// lightning.network/lightning-network-paper.pdf.

[22] Raiden Network. Available at: https://raiden.network.

[23] Bellare, M., &Rogaway, P. (2005). "Introduction to modern cryptography". UcsdCse, 207, 207.

[24] Elliptic Curve Integrated Encryption Scheme for secp256k1 in Python. Available at: https://gith ubcom/ecies/py.

[25] Dingledine, R., Mathewson, N., &Syverson, P. (2004). "Tor: The second-generation onion router". *Naval Research Lab Washington DC.*

[26] Albrecht, M., Grassi, L., Rechberger, C., Roy, A., &Tiessen, T. (2016). "MiMC: Efficient encryption and cryptographic hashing with minimal multiplicative complexity". *In International Conference on the Theory and Application of Cryptology and Information Security*. pp. 191-219.

[27] Groth, J. (2016). "On the size of pairing-based non-interactive arguments". *In Annual international conference on the theory and applications of cryptographic techniques*. pp. 305-326.

[28] Circom. Available at: https://docs.circom.io/.

[29] Snarkjs. Available at: https://github.com/iden3/snarkjs.

[30] Web3.py. Available at: https://web3py.readthedocs.io.

## APPENDIX A: DEFINITION OF SECURITY

A DMC scheme Π = (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw, VerifyTransaction) is secure if it satisfies transaction unlinkability and overdraft safety. For security definitions, we design two experiments which are employed based on a stateful DMC oracle $\mathcal{O}^{\mathrm{DMC}}$. The $\mathcal{O}^{\mathrm{DMC}}$ provides queries for adversary $\mathcal{A}$, these queries being interfaces for executing the algorithms defined in Π. The oracle is initialized by public parameters $pp$ and stores a transaction set TX, a set NCS including a list of a tuple $(note, cm, sn)$ and a set of accounts Acc. Below, we describe these queries made to the oracle $\mathcal{O}^{\mathrm{DMC}}$.

- $\mathcal{Q}$= (**CreateAccount**). The challenger $\mathcal{C}$ : (i) computes a key pair $(sk, pk)$ and an address $addr$; (ii) adds $addr$ into Acc; (iii) outputs $(addr, pk)$.
- $\mathcal{Q}$= (**Deposit**, $v$, $pk_A$). The challenger $\mathcal{C}$: (i) computes a tuple $(note, cm, sn)$ and a transaction $\text{tx}_{\text{Deposit}}$ by calling Deposit algorithm; (ii) adds $(note, cm, sn)$ to NCS and $\text{tx}_{\text{Deposit}}$ to TX.
- $\mathcal{Q}$ = (**OpenChannel**, $note_1, note_2, v_{AB}, pk_B, diff_{AB}, ddl_{AB}$ ). The challenger $\mathcal{C}$ : (i) computes two tuples $(chnt_{AB}, chcm_{AB}, chsn_{AB})$ and $(note_r, cm_r, sn_r)$ and a transaction $\text{tx}_{\text{OpenChannel}}$ by calling OpenChannel algorithm; (ii) adds these two tuples to NCS and $\text{tx}_{\text{OpenChannel}}$ to TX.
- $\mathcal{Q}$= (**CloseChannel**, $\text{tx}_{\text{OffchainTransfer}}, x$). The challenger $\mathcal{C}$: (i) computes $\text{tx}_{\text{CloseChannel}}$ by calling CloseChannelbyDiff algorithm; (ii) adds $\text{tx}_{\text{CloseChannel}}$ to TX. Note that we only consider the case that the recipient actively closes the channel.
- $\mathcal{Q}$ = (**Withdraw**, $v, addr$ ). The challenger $\mathcal{C}$ : (i) computes $\text{tx}_{\text{Withdraw}}$ by calling Withdraw algorithm; (ii) adds $\text{tx}_{\text{Withdraw}}$ to TX.
- $\mathcal{Q}$= (**Insert**, tx). The challenger $\mathcal{C}$ verifies the output of VerifyTransaction algorithm: if the output is 1, adds the tx to TX; otherwise, it aborts.

### A.1 Transaction Unlinkability

Let $\mathcal{T}$ be the set of transaction $\text{tx}_{\text{OffchainTransfer}}$ attached with **CloseChannel** queries. We define the transaction unlinkability experiment $\text{DMC}_{\Pi,\mathcal{A}}^{\text{TR-UL}}(\lambda)$as follows.

1) The public parameters $pp$ = Setup$(1^{\lambda})$ are computed and provided to $\mathcal{A}$.
2) Whenever $\mathcal{A}$ queries $\mathcal{O}^{\text{DMC}}$, answer this query with transaction set TX at each step.
3) Continue answering queries until $\mathcal{A}$ sends a pair of transactions $(\text{tx}, \text{tx}')$ with the requirements: (i) $(\text{tx}, \text{tx}' \in \mathcal{T})$; (ii) $\text{tx} \neq \text{tx}'$; (iii) the senders and recipients of $\text{tx}, \text{tx}'$ are not $\mathcal{A}$.
4) The experiment outputs 1 if the senders of $(\text{tx}, \text{tx}')$ are same and the recipients of $(\text{tx}, \text{tx}')$ are also same. Otherwise, it outputs 0.

**Definition 2** (TR-UL Security)**.** A DMC scheme $\Pi$ = (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw, VerifyTransaction, VerifyOffchainTransfer) is TR-UL secure, if for PPT adversary $\mathcal{A}$, there is a negligible function $negl$ such thatPr$[\text{DMC}_{\Pi,\mathcal{A}}^{\text{TR-UL}}(\lambda) = 1] \leq negl(\lambda)$.

### A.2 Overdraft Safety

We design the overdraft safety experiment, which means PPT adversary $\mathcal{A}$ tries to attack a given DMC scheme. Firstly, we define five variables for the security model.

- $v_{\text{Deposit}}$, the total value deposited by $\mathcal{A}$. To compute $v_{\text{Deposit}}$, the challenger $\mathcal{C}$ finds out all Deposit transactions recorded in TX via **Deposit** queries and sums up these values which were transferred from $\mathcal{A}$.

- $v_{\text{Acc}} \rightarrow \mathcal{A}$, the total value received by $\mathcal{A}$ from accounts in Acc. To compute $v_{\text{Acc}} \rightarrow \mathcal{A}$, the challenger $\mathcal{C}$ looks up all $\text{tx}_{\text{OffchainTransfer}}$ in CloseChannel transactions recorded in TX via **CloseChannel** queries and adds the values whose recipient are $\mathcal{A}$.

- $v_{\text{Withdraw}}$, the total value redeemed by $\mathcal{A}$. To compute $v_{\text{Withdraw}}$, the challenger $\mathcal{C}$ finds out all Withdraw transactions recorded in TX via **Withdraw** queries and sums up these values which were transferred to $\mathcal{A}$.

- $v_{\mathcal{A}} \rightarrow \text{Acc}$, To compute $v_{\mathcal{A}} \rightarrow \text{Acc}$, the challenger $\mathcal{C}$ looks up all $\text{tx}_{\text{OffchainTransfer}}$ in CloseChannel transactions recorded in TX via **CloseChannel** queries and adds the values whose sender are $\mathcal{A}$.

- $v_{\text{unspent}}$, the spendable amount in $cm$ and $chcm$. The challenger $\mathcal{C}$ can check whether corresponding $note/chnt$ is spendable as follows. For $cm$, $\mathcal{C}$ checks if a **Withdraw** query which redeems $note$ generates a valid transaction $\text{tx}_{\text{Withdraw}}$. For $chcm$, $\mathcal{C}$ first uses $chnt$ to create an off-chain transaction $\text{tx}_{\text{OffchainTransfer}}$ via a **OffchainTransfer** query, and then checks if a **CloseChannel** query yields a valid transaction $\text{tx}_{\text{CloseChannel}}$, which closes the channel $chnt$ using $\text{tx}_{\text{OffchainTransfer}}$.

For an honest account $u$, $v_{\text{Withdraw}} + v_{\mathcal{A}\rightarrow\text{Acc}} + v_{\text{unspent}} > v_{\text{Deposit}} + v_{\text{Acc}\rightarrow\mathcal{A}}$.

Formally, we define the overdraft safety experiment $\text{DMC}_{\Pi,\mathcal{A}}^{\text{OD-SF}}(\lambda)$ as follows.

1) The public parameters $pp$ = Setup $(1^{\lambda})$ are computed and provided to $\mathcal{A}$.
2) Whenever $\mathcal{A}$ queries $\mathcal{O}^{\text{DMC}}$, answer this query with transaction set TX at each step.
3) Continue answering queries until $\mathcal{A}$ sends a set NCS.
4) Compute the five variables mentioned above.
5) The experiment outputs 1 if $v_{\text{Withdraw}} + v_{\mathcal{A}\rightarrow\text{Acc}} + v_{\text{unspent}} > v_{\text{Deposit}} + v_{\text{Acc}\rightarrow\mathcal{A}}$. Otherwise, it outputs 0.

**Definition 2** (OD-SF Security)**.** A DMC scheme $\Pi$ = (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw,VerifyTransaction, VerifyOffchainTransfer) is OD-SFsecure, if for PPT adversary $\mathcal{A}$, there is a negligible function $negl$ such that $\Pr[\text{DMC}_{\Pi,\mathcal{A}}^{\text{OD-SF}}(\lambda) = 1] \leq negl(\lambda)$.

## APPENDIX B: PROOF OF SECURITY

A DMC scheme $\Pi$ = (Setup, Deposit, OpenChannel, OffchainTransfer, CloseChannel, Withdraw, VerifyTransaction, VerifyOffchainTransfer) is secure if it satisfies transaction unlinkability and overdraft safety.

### B.1 Proof of Transaction Unlinkability

Let $\mathcal{T}$ be the set of transaction $\text{tx}_{\text{OffchainTransfer}}$ attached with **CloseChannel** queries. $\mathcal{A}$ wins the TR-UL experiment when it outputs a pair of transactions $(\text{tx}, \text{tx}')$ if the senders of $(\text{tx}, \text{tx}')$ are same and the recipients of $(\text{tx}, \text{tx}')$ are also same. Suppose $\mathcal{A}$ outputs a pair of transactions $\text{tx}_{\text{CloseChannel}}, \text{tx}'_{\text{CloseChannel}}$. The $\text{tx}_{\text{OffchainTransfer}}$ in $\text{tx}_{\text{CloseChannel}}$ satisfies the following equations:

1) $\text{tx}_{\text{OffchainTransfer}} = (chsn_{AB}, cm_B, cm_A, rt, \pi_{\text{OffchainTransfer}})$.
2) $cm_B$ = $\text{Comm}(pk_B, v_B, r_B)$.
3) $cm_A$ = $\text{Comm}(pk_A, v_A, r_A)$.

and the $\text{tx}'_{\text{OffchainTransfer}}$ in $\text{tx}'_{\text{CloseChannel}}$ satisfies the following equations:

1) $tx'_{\text{OffchainTransfer}} = (chsn'_{AB}, cm'_{B}, cm'_{A}, rt', \pi'_{\text{OffchainTransfer}})$.
2) $cm'_{B}$ =Comm$(pk'_{B}, v'_{B}, r'_{B})$.
3) $cm'_{A}$ =Comm$(pk'_{A}, v'_{A}, r'_{A})$.

$\mathcal{A}$ wins the TR-UL experiment if the senders and recipients contained in ($\text{tx}_{\text{OffchainTransfer}}, \text{tx}'_{\text{OffchainTransfer}}$ are the same, i.e., $pk_A = pk'_A$ and $pk_B = pk'_B$. There are two ways for $\mathcal{A}$ to distinguish whether $pk_i = pk'_i, i \in \{A, B\}$: (i) distinguish public keys from commitments; (ii) distinguish public keys from the zero-knowledge proofs.

For condition (i), $\mathcal{A}$ must distinguish $pk_i = pk'_i$ based on different commitments $(cm_i, cm'_i), i \in \{A, B\}$ without knowing other secret values, which means that $\mathcal{A}$ ought to break the hiding property of the commitment scheme. For condition (ii), $\mathcal{A}$ must distinguish $pk_i = pk'_i, i \in \{A, B\}$ based on different zero-knowledge proofs $\pi_{\text{OffchainTransfer}}, \pi'_{\text{OffchainTransfer}}$, which means that $\mathcal{A}$ ought to break the proof of knowledge property of the zk-SNARKs. However, due to the security of commitment scheme and zk-SNARks, $\mathcal{A}$ cannot distinguish the two pairs of public keys.

### B.2 Proof of Overdraft Safety

We modify the overdraft safety experiment without affecting the view of $\mathcal{A}$. First, for each $\text{tx}_{\text{OffchainTransfer}}$ inside CloseChannel transaction $\text{tx}_{\text{CloseChannel}}$ in TX, $\mathcal{C}$ computes a witness $\vec{\omega} = (chnt_{AB}, note_B, note_A, sk_A, path_{AB})$ for the instance $\vec{x} = (chsn_{AB}, cm_B, cm_A, rt)$. Then, $\mathcal{C}$ constructs an augmented transaction set (TX, W), a list of matched pairs $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$.

**Definition 3** (Overdraft safe transaction set)**.** An augmented transaction set (TX, W) is overdraft safe if the following holds.

1) Each $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$ in (TX, W) contains openings (e.g., $chsn_{AB}$) of a channel note commitment $chcm_{AB}$, which is the output of a transaction $\text{tx}_{\text{OpenChannel}}$ that precedes $\text{tx}_{\text{OffchainTransfer}}$ in TX.

2) No two $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$ and $(tx'_{\text{OffchainTransfer}}, \vec{\omega}')$ in (TX, W) contain openings of the same note commitment.

3) Each $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$ in (TX, W) contains openings of $chcm_{AB}, cm_B, cm_A$ to values $v_{AB}, v_B, v_A$ respectively, satisfying $v_{AB} = v_B, +v_A$.
4) For each $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$ in (TX, W), if $chcm_{AB}$ is the output of a transaction $\text{tx}_{\text{OpenChannel}}$ in TX, then its witness $\omega$ contains an opening of $chcm_{AB}$ to a value $v$ that is equal to $v_{AB}$.

5) For each $(\text{tx}_{\text{OffchainTransfer}}, \vec{\omega})$ in (TX, W), where $\text{tx}_{\text{OpenChannel}}$ is inserted by $\mathcal{A}$, it holds that if $chcm_{AB}$ is the output of an earlier transaction $\text{tx}_{\text{OpenChannel}}$, then the public value $v$ in $\text{tx}_{\text{OpenChannel}}$ is equal to $chcm_{AB}$.

One can prove that (TX, W) is overdraft safe if the equation holds: $v_{\text{Withdraw}} + v_{\mathcal{A}\rightarrow\text{Acc}} + v_{\text{unspent}} > v_{\text{Deposit}} + v_{\text{Acc}\rightarrow\mathcal{A}}$. For each case mentioned above, we prove that five cases are in a negligible probability by way of contradiction. Note that we denote by $\Pr[\mathcal{A}(\overline{Con_k}) = 1]$ a non-negligible probability that $\mathcal{A}$ wins but violates condition $k, k \in \{1,2,3,4\}$.

**$\mathcal{A}$ violates Condition 1.** During construction of $\mathcal{O}^{\mathrm{DMC}}$, every $(\mathrm{tx}_{\mathrm{OffchainTransfer}}, \vec{\omega})$ in (TX, W) where $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ is not inserted by $\mathcal{A}$ satisfies condition 1; thus, $\Pr[\mathcal{A}(\overline{Con_1}) = 1]$ is a probability that $\mathcal{A}$ inserts $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ to construct $(\mathrm{tx}_{\mathrm{OffchainTransfer}}, \vec{\omega}) \in$ (TX, W) where $chcm_{AB}$ used in $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ is not the output note commitment of any previous transactions before $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ in TX.

Note that the validity of $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ implies that the witness $\omega$ contains a valid path $path_{AB}$ for a Merkle tree constructed by commitments in earlier transactions. However, a contradiction can be found: if $chcm_{AB}$ does not previously exist in TX, then $path_{AB}$ is not a valid path but with a valid root $rt$. Therefore, this violates the property of collision resistance of CRH.

**$\mathcal{A}$ violates Condition 2.** When condition 2 is violated, TX contains two transactions $\mathrm{tx}_{\mathrm{OffchainTransfer}}$ and $tx'_{\mathrm{OffchainTransfer}}$ that spend the same note commitment $chcm_{AB}$, and yield two different serial numbers $chsn_{AB}$ and $chsn'_{AB}$. Obviously, $\Pr[\mathcal{A}(\overline{Con_2}) = 1]$ is a probability that $\mathcal{A}$ inserts a pair of transactions where $chcm_{AB} = chcm'_{AB}$ and $chsn_{AB} \neq chsn'_{AB}$. However, if the two transactions spend the same $chcm_{AB}$ but create different serial numbers, then corresponding witnesses $\omega$ and $\omega'$ include different opening of $chcm_{AB}$.Therefore, this contradicts the binding property of the commitment scheme.

**$\mathcal{A}$ violates Condition 3.** $\Pr[\mathcal{A}(\overline{Con_3}) = 1]$ is a probability that the equation $v_{AB} \neq v_B + v_A$ holds. When violating condition 3, the equation $v_{AB} = v_B + v_A$ does not hold, so violating the soundness of zk-SNARKs during the construction of zero-knowledge proof $\pi_{\mathrm{OffchainTransfer}}$.

**$\mathcal{A}$ violates Condition 4.** Each $(\mathrm{tx}_{\mathrm{OffchainTransfer}}, \vec{\omega})$ in (TX, W) contains values (i.e., $v_{AB}$) of $chcm_{AB}$, and $chcm_{AB}$ is also the output commitment to values (including $v'_{AB}$) in a OpenChannel transaction $\mathrm{tx}_{\mathrm{OpenChannel}}$. Obviously, $\Pr[\mathcal{A}(\overline{Con_4}) = 1]$ is a probability that the euqation $v_{AB} \neq v'_{AB}$ holds. Thus, this contradicts the binding property of commitment scheme.

## AUTHORS

**Su Liu**, female, master's degree. Her main research fields are blockchain technology, privacy protection, etc

**Jian Wang**, male, doctor, professor, doctoral supervisor. His main research fields are key management, cryptographic protocol, privacy protection, etc.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# Creating multi-scripts sentiment analysis lexicons for Algerian, Moroccan and Tunisian dialects

K. Abidi and K. Smaïli

Loria - University Lorraine, France

**Abstract.** In this article, we tackle the issue of sentiment analysis in three Maghrebi dialects used in social networks. More precisely, we are interested by analysing sentiments in Algerian, Moroccan and Tunisian corpora. To do this, we built automatically three lexicons of sentiments, one for each dialect. Each lexicon is composed of words with their polarities, a dialect word could be written in Arabic or in Latin scripts. These lexicons may include French or English words as well as words in Arabic dialect and standard Arabic. The semantic orientation of a word represented by an embedding vector is determined automatically by calculating its distance with several embedding seed words. The embedding vectors are trained on three large corpora collected from YouTube. The proposed approach is evaluated by using few existing annotated corpora in Tunisian and Moroccan dialects. For the Algerian dialect, in addition to a small corpus we found in the literature, we collected and annotated one composed of 10k comments extracted from Youtube. This corpus represents a valuable resource which is proposed for free [1].

**Keywords:** Maghrebi dialect · Word embedding · Semantic orientation.

## 1 Introduction

To understand the requirements of users, clients or people in general, it is necessary to mine social media [3, 4, 30] and to develop automatic tools allowing a systematic analysis of the contents. One can then extract useful information that could be used in marketing advising, political views, movies reviews, etc. Henceforth, proposing methods to understand opinions is necessary and it is considered as a challenging issue especially for under-resourced languages such as Maghrebi dialects.

In this article, we will address the issue of developing a method allowing to analyse sentiments in the three following Maghrebi dialects: Algerian, Moroccan and Tunisian. The problem is that these dialects are under-resourced because they are not formal and not official languages. Basically, Arabic dialects are founded on Modern Standard Arabic (MSA), but not only. The originality of this work

[1] https://smart.loria.fr/corpora

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 147-156, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111413

is to propose a sentiment analysis tackling two issues frequent in Algerian, Moroccan and Tunisian dialects: the code-switching nature of a document and its multi-script form.
In order to explain the importance of this research work, let's give in the following the different particularities of the Maghrebi dialects.

The origin of Maghrebi dialects is mainly Standard Arabic, but not only. For practical reasons, several morpho-syntactic rules of MSA are not respected in Arabic dialects. This means that it is difficult to use the amount existing NLP resources developed for MSA to process Arabic dialects.

The vocabularies of the Arabic dialects evolve continuously by introducing new words, that could be considered as gibberish such as the word *papicha* that means *beautiful girl* in Algerian dialect. And as in any other language, Maghrebi dialects can borrow new words and adapt them phonologically to the local dialect such as: كوزِينة (borrowed from the French word *cuisine* that means *kitchen*).

Another particularity of Maghrebi dialects is the fact that people can write Arabic by using multiple scripts: Arabic and Latin [6, 7]. In addition, in social media people can use digits when they write in Latin script to represent sounds that do not exist in French or English, such as ع which is replaced by the digit 3.

In addition to these phenomena, in north Africa, code-switching is common in conversations. One can mix in the same sentence local Arabic, MSA, and foreign languages, such as French or English. In the following, we give an example:

تبَارك اللّه عليك *merci pour la vidéo* عجبني *les boucles d'oreilles* منين خديتِهم **thanks.**
**Translation** : "*God bless you, thank you for this video, I really liked the earrings where did you buy them. Thank you.*"

In this sentence, one switched from Moroccan dialect written in Arabic script, to French, then to Arabic, then again to French then once again to Arabic and finally to English.

Only few works addressing the issue of sentiment analysis in Maghrebi dialects do exist. But most of them have ignored the problems already cited and have concentrated on sentiment analysis in Maghrebi texts written only in Arabic script. In this article, we propose a method that allows to create automatically sentiment lexicons for Arabic dialects taking into account all the phenomena aspects related to Maghrebi dialects. This approach could be adapted to any Arabic dialect and also to any other low-resourced languages.

The rest of the paper is organized as follows: Section 2 is dedicated to the related work, while Section 3 examines the corpus we harvested. In Section 4, we discuss the proposed method to analyse sentiment of Maghrebi dialect. In Section 5, we present the different used corpora and the experimental results and finally we conclude.

## 2 Related work

Many studies have been conducted to address the issue of sentiment analysis in Arabic documents [2, 22, 8]. Researchers proposed various interesting approaches, that we can classify into two categories: machine learning techniques [10, 20, 12, 28, 18] and lexicon-based approaches [27, 15, 1, 14, 21]. Unfortunately, most of these methods are not directly reusable for Arabic dialects for the reasons mentioned in the introduction. In this section, we discuss the research works proposed to analyse sentiments in Arabic dialects and we will focusing on those concerning the three Maghrebi ones studied in this article.

In the Arabic dialect sentiment analysis literature, several works have used machine learning techniques to address this issue. These methods require a significant amount of pre-annotated corpora to train a good classifier that is able to distinguish between positive and negative documents.

In [16], the authors proposed a deep learning model based on Long short-term memory (LSTM) architecture to identify the sentiments of documents written in Egyptian and Emirati dialects. To train this model, the authors collected and annotated a corpus of 470k tweets. This model achieved an accuracy of 70% and 63% for Egyptian and Emirati, respectively. The authors of [9] proposed a model that combines LSTM with a convolutional neural network architectures. They used two existing annotated corpora extracted from Twitter to train the proposed model, which is composed of 10k and 2k of tweets written in Egyptian and Levantine, respectively. The method achieved an accuracy of more than 85%. Deep LSTM architecture has been also used by the authors of [23] to tackle the issue of sentiment analysis in Tunisian Dialects. The authors trained the model on a Tunisian corpus composed of 17k and the method achieved an accuracy of 90%.

Unlike the machine learning techniques, in the based-lexicon method, the global sentiment of a document is estimated by calculating the semantic orientation of the words appearing within the text. This approach requires the use of a lexicon of words with their polarities (positive and negative). In this approach, the need of sentiment lexicons is crucial for analysing documents in terms of opinions, that is why the authors in [17] created semi-automatically a sentiment lexicon, where 45% of the entries are Egyptian while 55% are words of Modern Standard Arabic. A sentiment lexicon for the Khaliji dialect has been built manually by exploring and labelling the words of a Saudi dialect twitter corpus (SDTC) [11]. The authors of [24] built a lexicon for Algerian dialect by translating manually an existing Egyptian polarity lexicon. In [13], the authors proposed an approach for emotion analysis of Tunisian comments posted in Facebook by using an emotion dictionary created automatically.

Actually, only a limited number of researches have been carried out for sentiment analysis in the three Maghrebi dialects, while a majority of research in this scope are dedicated to texts written only in Arabic script.

## 3 The collected dataset

In this work, we are interested by the Algerian, Moroccan and Tunisian dialects. That is why, we extracted three large corpora from YouTube by using the approach proposed in [5]. This method crawls (by using the API[2]) the posts of videos using specific hashtags related to each country.

In Table 1, we give some figures about the collected data, where $|C|$ indicates the number of comments, $|W|$ the number of words and finally $|V|$ the size of the vocabulary. We mention that these statistics concern the data obtained after the cleaning process.

Table 1: Statistics of the harvested corpora.

| | **Algerian (M)** | **Moroccan (M)** | **Tunisian(M)** |
|---|---|---|---|
| $\|C\|$ | 1.61 | 1.60 | 1.26 |
| $\|W\|$ | 23 | 22 | 17 |
| $\|V\|$ | 1.2 | 1.3 | 1 |

## 4 The sentiment analysis method

In this work, we propose a lexicon-based approach to analyse the sentiments of Maghrebi comments extracted from social networks. In this approach, the polarity of a text can be obtained on the ground of the polarity of the words that compose it. To do this, a lexicon of words, where each entry is associated to its polarity is necessary.

Because, in the Maghrebi dialects people use Latin and Arabic scripts and foreign languages to post their comments, we aim, in this work to handle this issue by building a multi-script and multilingual sentiment lexicon in order to analyze the sentiments of the collected corpus. A word and its polarity constitute an entry in the lexicon. Each word of this lexicon can be written in Arabic or Latin script and it can belong to one of the following languages: one of the three Maghrebi dialects, MSA, French or English as in the table 2. Concerning the polarity of each entry, we determined it automatically by using an approach similar to the one used in [29]. In this method, the authors proposed to tag the words by using the polarities of a small list of words called *seed words* for which the polarity is assigned by hand. A word is attached to the dominant polarity of the closest seed words. For example, a word is considered positive if it is closer, in terms of distance, to positive seed words than to negative words.

In the following, we will detail the two main steps of the method we used to assign a polarity to an entry of a sentiment lexicon.

[2] https://developers.google.com/YouTube

Table 2: Few examples concerning the different forms of a word

| **Script** | **language** | **Word** | **Meaning** |
|---|---|---|---|
| Latin | Algerian dialect | Chaba | Beautiful |
| Arabic | Algerian dialect | شَابة | Beautiful |
| Arabic | MSA | كَاذب | Liar |
| Latin | MSA | Kadib | Liar |
| Latin | English | Like | Like |
| Arabic | English | لَايك | Like |

## 4.1 Seed words identification

In [29], the authors used a list of seed words made up of seven positive words and seven negative words. In our case, for each dialect, we manually annotated a list of forty seed words for each of the polarities. The seed words have been selected, from a list of the most frequent not neutral words of the collected corpora. Then, we assigned manually to each of them its corresponding polarity. The number of seed words retained is relatively high compared to the experiment carried out by the authors of [29]. This is due to the fact that we want to cover as many words as possible since we deal with multi-script and multilingual words in our corpora. In the table 3 we give some examples of these retained seed words.

Table 3: Some examples of positive and negative seed words for the three dialects.

| **Algerian** | | **Moroccan** | | **Tunisian** | |
|---|---|---|---|---|---|
| **Positive** | **Negative** | **Positive** | **Negative** | **Positive** | **Negative** |
| chaba<br>(*Pretty*) | شيَات<br>(*Groveller*) | روعة<br>(*Wonderful*) | Problème<br>(*Problem*) | To9tool<br>(*so beautiful*) | مَاسط<br>(*Boring*) |
| Bravo | Mosakh<br>(*Disgusting*) | Hbiba<br>(*Sweetie*) | ينعل<br>(*Cursed*) | Ma7lek<br>(*How beautiful you are*) | جَاهل<br>(*Ignorant* ) |
| هَايل<br>(*Excellent*) | Na3ja<br>(*A weak personality* ) | Tbarklah<br>(*Marvelous* ) | Himar<br>(*Donkey*) | تهبل<br>(*Wonderful* ) | حيوَان<br>(*Beast* ) |
| الصحة<br>(*Health*) | Roukhs<br>(*Asshole* ) | كنحمَاق<br>(*I love*) | حرَام<br>(*Not good* ) | حلوه<br>(*Delicious* ) | حنش<br>(*Snake* ) |

## 4.2 Estimation of the polarity of words

We propose to calculate the semantic orientation of a word $w$ according to the difference between its closest positive seed words $SW_{pos}$ and its closest negative seed words $SW_{neg}$. We estimate the degree of closeness between two words one of which is a seed word by using the cosine similarity as in the formula we propose in 1:

$$SO(w) = \sum_{w_p \in SW_{pos}} Cos(w, w_p) - \sum_{w_n \in SW_{neg}} Cos(w, w_n) \qquad (1)$$

$w$ is considered as positive if $SO(w)$ is positive, similarly $w$ is considered as negative whether its orientation is negative. The similarity of the word $w$ and a seed word is estimated by using the cosine measure between their corresponding embedding vectors. The embedding vectors are generated by using the Continuous Bag-of-Words (CBOW), one of the method of the Word2Vec approach proposed by Mikolov [26]. The training has been achieved on the three large corpora presented in section 3. This led to the creation of a lexicon of sentiments for each of the dialects cited in this research. In the table 4, we detail each of them.

Table 4: Some figures about the lexicons of sentiments

| | **Algerian** | **Moroccan** | **Tunisian** |
|---|---|---|---|
| **Number of entries** | 11243 | 23405 | 10810 |
| **Positive words** | 8372 | 2326 | 19128 |
| **Negative words** | 2871 | 8484 | 4277 |

In table 5, we give some examples extracted from our lexicons.

Table 5: Some examples of positive and negative words extracted from the three lexicon

| **Word** | **Translation** | **Polarity** | **Lexicon** |
|---|---|---|---|
| ضحكاتنى | She makes me laughing | Positive | Moroccan |
| Kanbghiwk | We like you | Positive | Moroccan |
| البرهوش | Stubborn | Negative | Moroccan |
| frahnalk | Happy for you | Positive | Algerian |
| حركي | Groveller | Negative | Algerian |
| wetek | It suits you very well | Positive | Tunisian |
| ثعالب | Crafty | Negative | Tunisian |

## 5 Experimentation

We used the created lexicons to evaluate the polarity of the four following labeled corpora.

- **ElecMorocco**: is a Moroccan corpus extracted from Facebook and annotated by the authors of [19]. The main topic of this corpus concerns the local elections. It is constituted by 6389 positive and 4367 negative comments. This corpus contains only comments written in Arabic script.

- **TSAC**: is a Tunisian corpus collected from comments posted on official Facebook pages of Tunisian radios and TV channels [25]. It is composed of 5081 positive and 6514 negative comments written in Arabic and Latin scripts.

- **CorpusAlg**: is an Algerian corpus extracted from different Facebook pages, it contains 5079 posts, among them 3032 are positive comments [24]. The comments in this corpus are written in Latin and Arabic characters.

- **SentAlg**: The previous Algerian corpus (CorpusAlg) is small in comparison to the two others, that is why we decided to collect and annotate manually 10k of comments extracted from Algerian YouTube comments. This achieved a corpus of positive and negative comments of 5562 and 4438 respectively. All of the comments are written in Arabic and Latin script.

  Table 6 summarizes the figures of the different corpora used in our experimentation.

Table 6: Some figures about the different corpora.

| | **ElecMorocco** | **TSAC** | **CorpusAlg** | **SentAlg** |
|---|---|---|---|---|
| **Positive comments** | 3523 | 5081 | 3032 | 5562 |
| **Negative comments** | 6431 | 6514 | 2047 | 4438 |

### 5.1 Results

As mentioned before, we used in this work a sentiment analysis method which is based on sentiment lexicons and annotated corpora. In this method, the aim is to identify in the sentence under analysis, the words that exist in the lexicon and to take into account the corresponding polarities in the opinion to assign to the sentence. Then the evaluation is estimated by using scores such as: accuracy, recall, precision, F-measure, etc. In table 7, we give the achieved results in terms of Recall and Precision on the corpora listed above.

Table 7: Experimental results on the three Maghrebi corpus.

| | **Recall (%)** | **Precision(%)** | **F-measure (%)** |
|---|---|---|---|
| **ElecMorocco** | 59.29 | 63.23 | 61.19 |
| **TSAC** | 64.03 | 63.78 | 63.90 |
| **CorpusAlg** | 67.92 | 67.15 | 67.53 |
| **SentAlg** | **79.78** | **80.79** | **80.28** |

The results show that the weakest performance concerns the Moroccan corpus. An analysis of this corpus shown that this later contains a lot of sentences in Modern Standard Arabic, while our training was done on a corpus extracted from the comments of Youtube posted by Moroccan mostly in their dialect [6]. Another explanation of these results is due to the fact that ElectMoroccan contains only comments written in Arabic script which is not the case of the training corpus. Concerning the Tunisian and the Algerian test corpora TSAC and CorpusAlg, respectively, thematically, they are far from the crawled corpora used for the training. That is why the performances are reasonable, but they are not as good as the results achieved on the corpus SentAlg. We recall, that we used a lexicon-based approach for sentiment analysis, but unfortunately we have not found any available sentiment lexicon for the three studied dialects, that is why we created them automatically. In the opposite, SentAlg is a corpus which is similar thematically to the training corpus, which explains why the performances are higher than those obtained for the others.

## 6 Conclusion

We proposed, in this article, a lexicon-based approach to analyse sentiments of three Maghrebi dialects, namely Algerian, Moroccan and Tunisian. The dialects lexicons used to classify the documents in terms of sentiments were created automatically. The approach, we proposed depends on a predefined list of 80 polarity seed words for each dialect, selected manually. Then, using a similarity measure, we estimated the proximity between an embedding word and the list of the embedding seed words.

One of the originality of this method is that it allowed to create multi-script and a multi-lingual sentiment lexicons for Algerian, Moroccan and Tunisian dialects which contain 11.2k, 23.4k and 10.8k entries, respectively. These sentiment lexicons were used to classify, in terms of polarities, three test datasets. This approach has been also tested on an Algerian corpus we collected and labelled manually. The performances we achieved depend on the quality of the corpora and they vary between 61.99 and 80.28 in terms of F-measure.

## References

1. Abd-Elhamid, L., Elzanfaly, D., Eldin, A.S.: Feature-based sentiment analysis in online arabic reviews. In: 2016 11th International Conference on Computer Engineering Systems (ICCES) (2016)
2. Abdul-Mageed, M., Diab, M., Korayem, M.: Subjectivity and sentiment analysis of modern standard Arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011)
3. Abdul-Mageed, M., Diab, M.T., Kübler, S.: SAMAR: subjectivity and sentiment analysis for arabic social media. vol. 28, pp. 20–37 (2014)

4. Abidi, K., Fohr, D., Jouvet, D., Langlois, D., Mella, O., Smaïli, K.: A Fine-grained Multilingual Analysis Based on the Appraisal Theory: Application to Arabic and English Videos. vol. Communications in Computer and Information Science book series (CCIS, volume 1108), pp. 49–61. Springer, Nancy, France (2019)
5. Abidi, K., Menacer, M.A., Smaili, K.: CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube. In: 18th Annual Conference of the International Communication Association (Interspeech). Conference of the International Communication Association (Interspeech), Stockholm, Sweden (2017)
6. Abidi, K., Smaïli, K.: An empirical study of the Algerian dialect of Social network. In: ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing. Casablanca, Morocco (2017), https://hal.inria.fr/hal-01659997
7. Abidi, K., Smaïli, K.: An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings. In: 11th edition of the Language Resources and Evaluation Conference, LREC 2018. Miyazaki, Japan (2018), https://hal.archives-ouvertes.fr/hal-01718110
8. Abo, M.E.M., Raj, R.G., Qazi, A.: A review on arabic sentiment analysis: State-of-the-art, taxonomy and open research challenges (2019)
9. Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., Dobnik, S.: Lstm-cnn deep learning model for sentiment analysis of dialectal arabic. In: Smaïli, K. (ed.) Arabic Language Processing: From Theory to Practice (2019)
10. Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El Hajj, W., Bashir Shaban, K.: Deep learning models for sentiment analysis in Arabic. In: Proceedings of the Second Workshop on Arabic Natural Language Processing (Jul 2015)
11. Al-Thubaity, A., Alqahtani, Q., Aljandal, A.: Sentiment lexicon for sentiment analysis of saudi dialect tweets. vol. 142, pp. 301–307 (2018), arabic Computational Linguistics
12. Ali, A.E., Stratmann, T.C., Park, S., Schöning, J., Heuten, W., Boll, S.: Measuring, understanding, and classifying news media sympathy on twitter after crisis events. vol. abs/1801.05802 (2018)
13. Ameur, H., Jamoussi, S., Ben Hamadou, A.: Exploiting emoticons to generate emotional dictionaries from facebook pages. In: Czarnowski, I., Caballero, A.M., Howlett, R.J., Jain, L.C. (eds.) Intelligent Decision Technologies 2016 (2016)
14. Awwad, H., Alpkocak, A.: Performance comparison of different lexicons for sentiment analysis in arabic. In: 2016 Third European Network Intelligence Conference (ENIC) (2016)
15. Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W.: A large scale arabic sentiment lexicon for arabic opinion mining. In: ANLP@EMNLP (2014)
16. Baly, R., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., Shaban, K.B., El-Hajj, W.: Comparative evaluation of sentiment analysis methods across arabic dialects. Procedia Computer Science **117**, 266–273 (2017), arabic Computational Linguistics
17. El-Beltagy, S.: Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic (05 2016)
18. Elfaik, H., Nfaoui, E.H.: Deep bidirectional lstm network learning-based sentiment analysis for arabic text. vol. 30, pp. 395–412 (2021)
19. Elouardighi, A., Maghfour, M., Hammia, H., Aazi, F.Z.: Analyse des sentiments à partir des commentaires facebook publiés en arabe standard ou dialectal marocain par une approche d'apprentissage automatique. In: Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018. pp. 329–334 (2018)
20. Heikal, M., Torki, M., El-Makky, N.: Sentiment analysis of arabic tweets using deep learning. vol. 142, pp. 114–122 (2018), arabic Computational Linguistics

21. Htait, A., Fournier, S., Bellot, P.: Identification semi-automatique de mots-germes pour l'analyse de sentiments et son intensité. In: COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings. pp. 415–424 (2017)
22. Ibrahim, H.S., Abdou, S., Gheith, M.: Sentiment analysis for modern standard arabic and colloquial. vol. abs/1505.03105 (2015)
23. Jerbi, M.A., Achour, H., Souissi, E.: Sentiment analysis of code-switched tunisian dialect: Exploring rnn-based techniques. In: Arabic Language Processing: From Theory to Practice - 7th International Conference, ICALP 2019, Nancy, France, October 16-17, 2019, Proceedings. pp. 122–131 (2019)
24. Mataoui, M., Zelmati, O., Boumechache, M.: A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. vol. 110, pp. 55–70 (2016)
25. Medhaffar, S., Bougares, F., Estève, Y., Belguith, L.H.: Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In: Proceedings of the Third Arabic Natural Language Processing Workshop, WANLP 2017@EACL, Valencia, Spain, April 3, 2017. pp. 55–61 (2017)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (Workshop) (2013)
27. Mohammad, S., Dunne, C., Dorr, B.J.: Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: EMNLP (2009)
28. Nejjari, M., Meziane, A.: Sahar-lstm: An enhanced model for sentiment analysis of hotels'arabic reviews based on lstm (2020)
29. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association (2003)
30. Yimam, S.M., Alemayehu, H.M., Ayele, A., Biemann, C.: Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In: Proceedings of the 28th International Conference on Computational Linguistics (Dec 2020)

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN INTELLIGENT AND INTERACTIVE GAMING SYSTEM TO PROMOTE ENVIRONMENT AWARENESS USING CONTEXT-BASED STORYING

Yilin Luo[1] and Yu Sun[2]

[1]Santa Margarita Catholic High School, Rancho Santa Margarita, CA 92688
[2]California State Polytechnic University, Pomona, CA, 91768

## ABSTRACT

*Since a child, I loved to play video games, especially platform games such as Metal Slug™, Mega Man™, etc.. Therefore, I was inspired to design my own platform game; with this paper, I have the opportunity to introduce our platform game, which is a "JFF Game" we developed using Unity and Visual Studio 2019.*

## KEYWORDS

*Machine Learning, 3D design, Gaming System.*

## 1. INTRODUCTION

I have loved to play platform games since a child and have learned a lot from doing so. For example, Metal Slug™ taught me that I should never give up when encountering difficulties, while Mega Man™ taught me that I should be honest all the time. I have always wanted to design my own platform game to express myself, and the development of our JFF Game has enabled me to do that. [1, 2, 3]

There are many software applications that can be used to make games, but each one will profoundly affect the way it is developed. There are two well-known tools in the field of game development. First, there is Unreal Engine. [4] The advantage of Unreal Engine is that the developer usage rate is high and it has many tools; however, some tools are difficult to learn and use. Second, there is CryEngine 3. [5, 6] The advantage of CryEngine 3 is that people can easily create complex and diversified special effects with it, but it also has disadvantages: for example, its developer community is not strong enough, and it is also difficult to learn.

We used Unity [7, 8, 9] and Visual Studio 2019 [10, 11, 12] to design our JFF Game. Compared with other software, the advantage of Unity is that it is easy to use and compatible with all systems (Windows, Mac, etc.). However, one issue with Unity and Visual Studio 2019 is that its available tools are limited, so it takes a long time to create complex and diversified effects. This was not so much of an issue for us, however, since our JFF Game does not require complex or diversified effects.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 157-162, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111414

We employed several methods to test our JFF Game. For example, we asked friends to try it to see if they encountered any bugs. After some modifications, we believe there are no remaining bugs in the current iteration of our JFF Game.

The rest of this paper is organized as follows: in section 2, we introduce the challenges we encountered while creating our JFF Game. In section 3, we introduce how the game works in detail. In section 4, we illustrate two experiments we conducted to design the game. In section 5, we introduce three related works. In the final section, we summarize the process of creating our JFF Game.

## 2. CHALLENGES

In order to design our JFF Game, which we developed using Unity and Visual Studio 2019, a few challenges were identified as follows.

### 2.1. Challenge 1: Choosing an overall context and main character

Our first challenge was to choose an overall context for the game. We wanted to make a platform game, but weren't sure which story frame might maximize the characteristics of this type of game and its perimeters. After watching a lot of videos and holding discussions with our team, we finally decided to use the medieval knight as our hero character. This figure appears in many classic works of Literature, so we were eager to use such a character to express our love and respect for classical works.

### 2.2. Challenge 2: Designing the game map

Our second challenge was to design the game map. Since I wasn't sure how to design interesting maps, we ended up designing several test maps and asking friends for feedback. After summarizing our friends' opinions, we were finally able to build a solid game map.

### 2.3. Challenge 3: Finding a useable plot line

Our third challenge was to design plots. We initially wanted to make a "prince saves the princess" story, yet we found this type of story unpopular. My mother made the suggestion that "You can think about this story in another way," and after listening to her, I suddenly had the idea that the importance of the prince's sword to the prince is similar to the importance of the princess to the prince. Therefore, we decided to write the story of a "prince looking for his sword."

## 3. SOLUTION

Our JFF Game is a platform game made by Unity and Visual Studio 2019. Players can use "W,S,A,D" to control the direction of the game character and "Space" to make the character jump. There are three levels in our game: in level one, the goal is to find the sword; in levels two and three, the goal is to escape. There are two types of enemies. The first enemy attacks overtly by cannon fire, using bullets that are fired to attack the player. The second enemy attacks stealthily by ambushing the player. The total HP of the player is 100, and the damage caused by these enemies varies with the type of attack suffered.

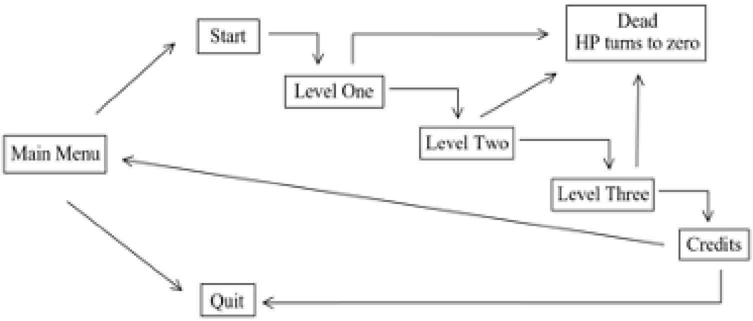
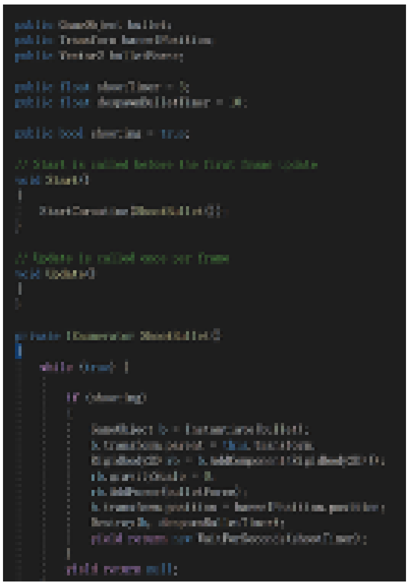


Figure 1. Game map

This segment of code (see Figure 2) describes how the overt type of enemy works. For example, from this segment of code, we can see that the shoot time is 5 and the despaen bullet time is 10.



```
public GameObject bullet;
public Transform barrelPosition;
public Vector2 bulletForce;

public float shootTimer = 5;
public float despawnBulletTimer = 10;

public bool shooting = true;

// Start is called before the first frame update
void Start()
{
    StartCoroutine(ShootBullet());
}

// Update is called once per frame
void Update()
{
}

private IEnumerator ShootBullet()
{
    while (true) {
        if (shooting)
        {
            GameObject b = Instantiate(bullet);
            b.transform.parent = this.transform;
            Rigidbody2D rb = b.AddComponent<Rigidbody2D>();
            rb.gravityScale = 0;
            rb.AddForce(bulletForce);
            b.transform.position = barrelPosition.position;
            Destroy(b, despawnBulletTimer);
            yield return new WaitForSeconds(shootTimer);
        }
        yield return null;
```

Figure 2. Code governing enemy behaviour

Below are screenshots from levels one and two (Figures 3, 4).

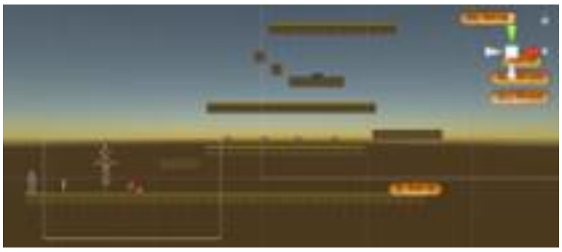


Figure 3. Screenshot from our JFF Game

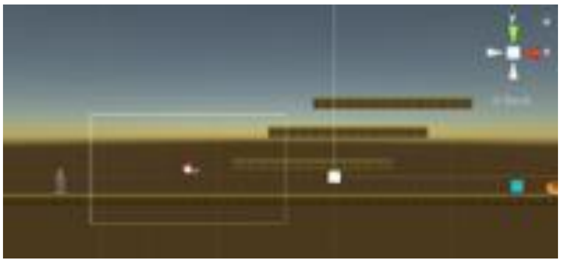


Figure 4 Screenshot from our JFF Game

When players touch the statue at the end of each level, they progress to next level automatically.

## 4. Experiment

We wanted to add a score system and time limit to our JFF Game, so we made a small game and invited ten friends to play it. Afterwards, we asked them to provide feedback. However, the results were surprising since many thought the scoring system and time limit both limited their ability to explore. Since players would likely only choose the pathway that would achieve the highest score possible instead of exploring others, we decided not to include a score system or time limit.

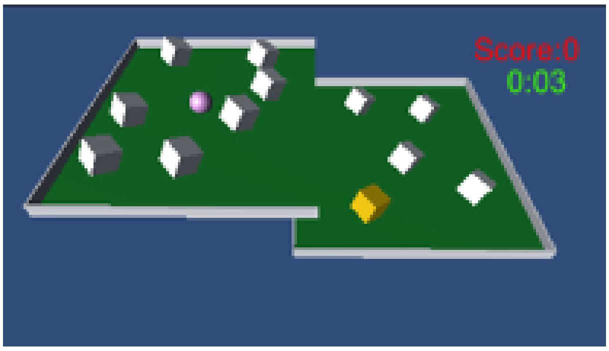
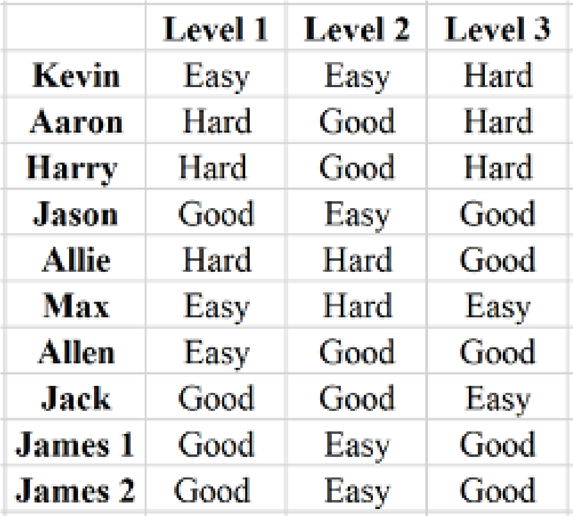


Figure 5. Screenshot of the test game

To test the difficulty of the game, we invited ten friends to play. Afterwards, we asked them to provide feedback. Our benchmark for editing was: more than three people saying it was “Good,” more than one person saying it was “Easy,” and more than one person saying it was “Hard.” If these variable criteria were present, then the design of the level was considered good and reasonable. For level one, 3 people said it was “Easy,” 3 people said it was “Hard,” and 4 people said it was “Good.” For level two, 4 people said it was “Easy”, 2 people said it was “Hard,” and 4 people said it was “Good.” For level three, 2 people said it was “Easy,” 3 people said it was “Hard,” and 5 people said it was “Good.” Therefore, the overall difficulty of our game was deemed appropriate (see figure 6).



| | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| Kevin | Easy | Easy | Hard |
| Aaron | Hard | Good | Hard |
| Harry | Hard | Good | Hard |
| Jason | Good | Easy | Good |
| Allie | Hard | Hard | Good |
| Max | Easy | Hard | Easy |
| Allen | Easy | Good | Good |
| Jack | Good | Good | Easy |
| James 1 | Good | Easy | Good |
| James 2 | Good | Easy | Good |

Figure 6. Feedback for game difficulty

Our experiments proved useful since experiment one allowed us to reconsider using a score system or time limit, and experiment 2 revealed that the difficulty of our game was on par with our expectations.

## 5. RELATED WORK

Burke, Quinn, and Yasmin B. Kafai introduce the tools and communities of game making. This source helped us realize that we could use more tools to complete our game. For example, the RPG maker is useful, and can help someone complete a game with rich elements without the use of programming. [13]

James Paul Gee explores how video games can also be learning machines. This work influenced us to belive we can use games in many ways, for different purposes; they're not just for making people happy, but also possibly educating them about something. [14]

Susan Bermingham, et al. discuss how to collaborate to make games. This essay influenced our belief that if we wanted to make a better and more interesting game, a key component of the design phase would be to collaborate with others and discuss ideas freely.

## 6. CONCLUSION AND FUTURE WORK

We used Unity and Visual Studio 2019 to make our JFF Game. We chose Unity because, compared to other software, it is easy to use and compatible with all systems. In our game, players use "W,S,A,D" to control the direction of the character and "Space" to make the character jump. Our JFF game is a platform game with three levels, and the ultimate goal is to find the sword and escape the forest. There are two kinds of enemies to avoid, and players have 100 HP.

We did two experiments. Before starting, we ran experiment 1. As a result, we deleted the score system and time limit that were initially present. After finishing the draft of our game, we ran experiment 2. Experiment two confirmed that we didn't need to change the difficulty level of our game.

There are still a few issues remaining to resolve. Our JFF Game has only three levels, so it may not be complex enough to satisfy some players. Also, there are no complex and diversified special effects, and the plot can be dull. The character's movements are still slightly stiff, and there are only two kinds of enemies, which may be boring for some players. Also, when players become familiar with the enemies' attack routines, the game can become too simple to remain challenging.

To resolve these issues, we may make the game more complex with more diversified special effects. To do this, we plan to collaborate with other people so the game can become better through continued input and testing.

## REFERENCES

[1] Schrier, Karen. "Designing and using games to teach ethics and ethical thinking." *Learning, education and games* 141 (2014).

[2] Shaffer, David Williamson, and James Paul Gee. *How computer games help children learn*. New York: Palgrave Macmillan, 2006.

[3] Shaffer, David Williamson, and James Paul Gee. *How computer games help children learn*. New York: Palgrave Macmillan, 2006.ß

[4] Sanders, Andrew. *An introduction to Unreal engine 4*. CRC Press, 2016.

[5] Tracy, Sean, and Paul Reindell. *CryENGINE 3 Game Development: Beginner's Guide*. Packt Publishing Ltd, 2012.

[6] Sousa, Tiago, Nickolay Kasyan, and Nicolas Schulz. "CryENGINE 3: Three Years of Work in Review." *GPU Pro* 3 (2012): 133-168.

[7] Halpern, Jared. "Introduction to unity." *Developing 2D Games with Unity*. Apress, Berkeley, CA, 2019. 13-30.

[8] Halpern, Jared, and Halpern. *Developing 2D Games with Unity*. New York City: Apress, 2019.

[9] Meijers, Alexander. "Unity." *Immersive Office 365*. Apress, Berkeley, CA, 2020. 87-119.

[10] Strauss, Dirk. "Getting to Know Visual Studio 2019." *Getting Started with Visual Studio 2019*. Apress, Berkeley, CA, 2020. 1-60.

[11] Strauss, Dirk. "Working with Visual Studio 2019." *Getting Started with Visual Studio 2019*. Apress, Berkeley, CA, 2020. 61-122.

[12] Strauss, Dirk. "Getting Started with Visual Studio 2019."

[13] Burke, Quinn, and Yasmin B. Kafai. "Decade of game making for learning: From tools to communities." *Handbook of digital games* (2014): 689-709.

[14] Gee, James Paul. "Learning by design: Good video games as learning machines." *E-learning and Digital Media* 2.1 (2005): 5-16.

[15] Bermingham, Susan, et al. "Approaches to collaborative game-making for fostering 21st century skills." *European Conference on Games Based Learning*. Academic Conferences International Limited, 2013.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# SENTIMENT ANALYSIS OF COVID-19 VACCINE RESPONSES IN MEXICO

Jessica Salinas, Carlos Flores, Hector Ceballos and Francisco Cantu

School of Engineering and Sciences,
Tecnologico de Monterrey, Monterrey, Mexico

### *ABSTRACT*

*The amount of information that social networks can shed on a certain topic is exponential compared to conventional methods. As new COVID-19 vaccines are approved by COFEPRIS in Mexico, society is acting differently by showing approval or rejection of some of these vaccines on social networks. Data analytics has opened the possibility to process, explore, and analyze a large amount of information that comes from social networks and evaluate people's sentiments towards a specific topic. In this analysis, we present a Sentiment Analysis of tweets related to COVID-19 vaccines in Mexico. The study involves the exploration of Twitter data to evaluate if there are preferences between the different vaccines available in Mexico and what patterns and behaviors can be observed in the community based on their reactions and opinions. This research will help to provide a first understanding of people's opinions about the available vaccines and how these opinions are built to identify and avoid possible misinformation sources.*

### *KEYWORDS*

*Twitter, Data Mining, Sentiment Analysis, Machine Learning, COVID-19.*

## 1. INTRODUCTION

Ever since COVID-19 was declared a global pandemic by the World Health Organization (WHO), different institutions and laboratories around the world have started a race against time to develop a vaccine with the highest possible efficacy. After a year of pandemic, the first vaccines are now available, and the vaccination process has started throughout the globe. However, this process has been affected by the opinions of people regarding the different vaccines [1].

Even though clinical studies have shown the efficacy of the vaccines against COVID-19, many people are not convinced of vaccination and have avoided this process. This behavior has become so dangerous, that the World Health Organization has included vaccine hesitancy in its top 10 global health threats in 2019 [2].

However, COVID-19 is not the only pandemic in which this phenomenon has occurred. When the H1N1 virus was first detected in the United States and spread across the world in 2009, Twitter was full of speculations and conspiracy theories about the origin of the virus and the development of vaccines. Given the influence of social networks on people's opinions and the effect on communication, previous studies have made use of Twitter data to provide an analysis of sentiments toward a specific topic or situation. For instance, the tweets developed during the H1N1 epidemic helped in the search for sentiments such as humor, sarcasm, frustration, relief, misinformation, and more [3]. Also, a similar situation was presented in the study by Bessi et al.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 163-178, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111415

[4], where it was found that friends played a major role in the exposure of false information when Ebola appeared at the end of 2013.

Regarding the current pandemic, recent sentiment analyses that have been performed have dealt with topics such as anxiety and panic [5], lack of support for isolation [6], prediction of self-awareness of precautionary procedures, forecasting of stock sectors [7], and public attitude towards the pandemic, among others. With the given data about tweets, data mining techniques can be applied to processed data to discover patterns among data, build networks, form clusters, and perform classification for making predictions that can provide valuable information about people and their behavior towards the current pandemic.

Given the tremendous effect of social networks on people's opinions and behaviors, this research aims to use data mining techniques to identify the sentiments of the Mexican population towards the different COVID-19 vaccines available in Mexico using Twitter data. The results of this study will provide an insight into the general opinion of the Mexican population regarding these vaccines, and the possibility to identify potential misinformation sources and concerns.

The overview of the proposed methodology is show in Figure 1 and its main contributions are:

- To provide a framework through which the general response of the Mexican population towards the different vaccines available in the countrycan be deciphered.
- To report insights and findings within Twitter data about the behavior of the Mexican population regarding the available vaccines in the country.
- To find and report different topics within the COVID-19 vaccine's conversations of Mexicans on Twitter using novel text mining approaches.
- To identify the main contributors/user profiles on the Mexican population on Twitter regarding the different COVID-19 vaccines using Social Network Analysis.

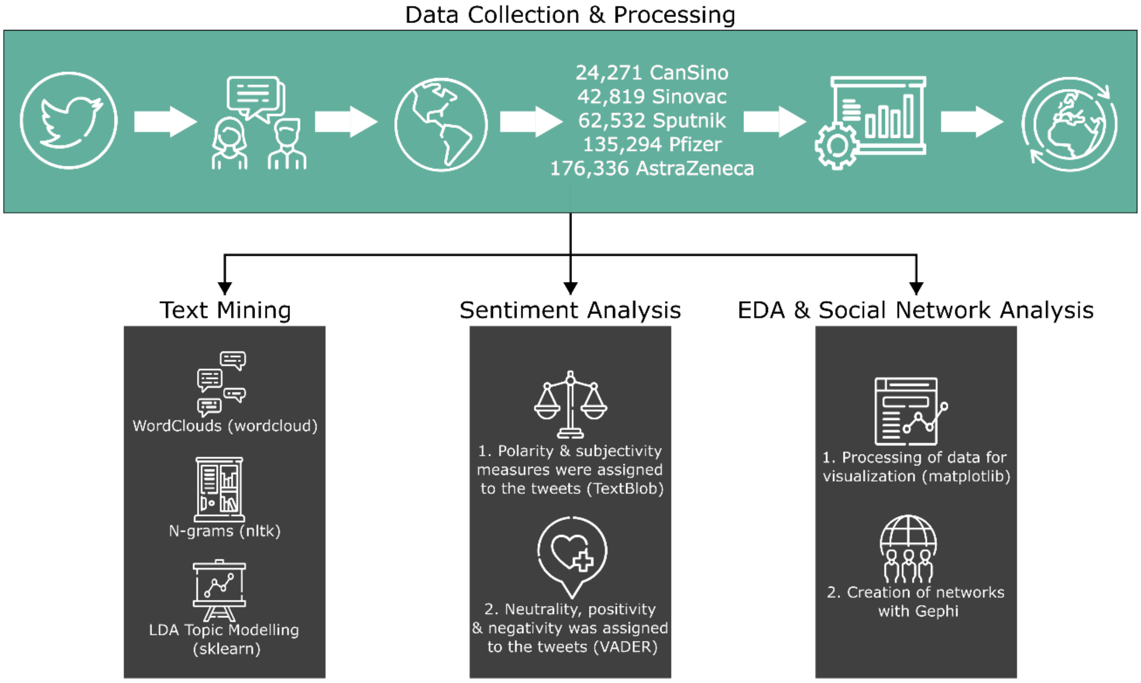

Data Collection & Processing
24,271 CanSino
42,819 Sinovac
62,532 Sputnik
135,294 Pfizer
176,336 AstraZeneca
Text Mining
WordClouds (wordcloud)
N-grams (nltk)
LDA Topic Modelling (sklearn)
Sentiment Analysis
1. Polarity & subjectivity measures were assigned to the tweets (TextBlob)
2. Neutrality, positivity & negativity was assigned to the tweets (VADER)
EDA & Social Network Analysis
1. Processing of data for visualization (matplotlib)
2. Creation of networks with Gephi

Figure 1. Proposed methodology.

This work is presented as follows: Section 2 presents related work with focus on sentiment classification on social media posts, studies related to Twitter data and the COVID-19 disease. Section 3 describes the methods used to carry out this work, focusing on the main tasks carried out such as data collection, data preparation, sentiment analysis, exploratory data analysis, and social network analysis. In Section 4, all the results obtained using modeling and analysis of data are presented, while in Section 5 the importance of previously stated results are explained and evaluated. Finally, Section 6 is a brief conclusion and summary of all the findings made.

## 2. BACKGROUND

Social networks have enabled people to express themselves and show their interest or disinterest over a wide range of topics and situations. such as specific brands, campaigns, or products. For instance, in the case of companies that involve sales, the nature of customers' comments or reviews that are published within social networks greatly affects how current and potential customers evaluate the company's products and decide whether to stay on the company or not. Furthermore, by making use of social networks' data, companies may identify these opinions and make decisions based on them to improve in target areas. Such approaches have been made between apparel brands [8], in the automotive industry [9], and in food chains [10], among others.

Data from social networks has also had several applications in sociological and psychological studies. Some of these applications include the application of sentiment analysis to study social and cultural aspects ranging from attitudes towards women [11], suicide [12], and literary studies [13], among others.

The current COVID-19 pandemic has also been the subject of studies that involve people's opinions and their interpretation. For instance, a study by Kumar et al. [14] focuses on analyzing public opinion about online learning during confinement due to the COVID-19 pandemic. Another article published by Toeppe et al. [15], presents an approach to evaluate the different emotions reflected among affective publics towards the current pandemic. For this study, data from Twitter was employed for the sentiment analysis. Similarly, Twitter data has been used to evaluate sentiments about the vaccines that have been developed against SARS-CoV-2. For example, in one of the studies [16], Twitter data was employed to evaluate the opinion of the Indonesian population about their vaccination scheme using tweets from January 2021. In another study [17], millions of tweets in English from February 2020 to December 2020 were analyzed to identify the emotions of the population on the upcoming vaccines for COVID-19.

While previous studies effectively use Twitter data for providing an insight into the response towards the different COVID-19 vaccines, vaccine-related tweets in Spanish have not been yet analyzed and the sentiments of the Mexican population towards the vaccines that are available within the country have not been explored either. For this reason, we present this study which focuses on the previous topic within the Mexican population.

## 3. METHODS AND DATA

### 3.1. Data Collection

The data used for this study was obtained using the Twitter Developer API. Tweets collected were filtered using the coordinates and radius of Mexico to obtain only the tweets from Mexican users. The vaccines selected for analysis were those acquired by the Mexican government to initiate the vaccination program in Mexico, which are: Pfizer-BioNTech, Sputnik, OxfordAstraZeneca, CanSino, and Sinovac. An individual dataset was created for each of the vaccines, initiating the collection of tweets from March 07, 2021, until April 21, 2021.

A total of 176,336, 135,294, 42,819, 24,271, and 62,532 tweets were gathered for the OxfordAstraZeneca, Pfizer-BioNTech, Sinovac, CanSino, and Sputnik vaccines, respectively. These tweets were then loaded into a Python Pandas data frame for further analysis.

### 3.2. Data Preparation

Once loaded into data frames, the tweets were cleaned by removing the duplicate tweets occasioned by re-tweets (RT), the hashtags, users, and unnecessary spaces in tweets. Afterwards, the tweets were translated to English to proceed with the Sentiment Analysis. This was carried out using the Googletrans Python library. After processing the data, the remaining number of tweets were of 31,905, 26,641, 7,511, 4,204, and 10,928 for Oxford-AstraZeneca, Pfizer-BioNTech, Sinovac, CanSino, and Sputnik vaccines, respectively.

### 3.3. Sentiment Analysis

After the datasets were processed, subjectivity and polarity measures were assigned to the tweets. This was carried out using the TextBlob Python library, using the Sentiment property.

Neutrality, positivity, and negativity measures were assigned to the tweets using the Valence Aware Dictionary for sEntiment Reasoning (VADER) [18] tool from the Natural Language Toolkit (NLTK) package in Python.

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a sentiment lexicon-based sentiment analyst constructed to work with text/sentiment expressed in social media microblogs. Its lexicon includes common features to sentiment expression in English such as emoticons (":)" referring to a smiley face, or ":(" referring to a sad face), acronyms ("LOL" and "WTF"), and slang ("meh", "nah"). Therefore, VADER outperforms other lexicon-based approaches in social media sentiment analysis. For instance, the approaches based on polarity lexicons (LIWC, GI, ...) are not able to capture the sentiment from emotions or acronyms since they are only capable of generating binary polarity. Although this polarity problem is solved by using valence-based sentiment approaches such as ANEW, this approach also fails to cover for most common lexical features in social media. Also, since VADER is not a machine learning model, problems such as the requirement of large datasets, high computational cost derived from training/classification time, and having features that are not easily interpretable due to processing in black boxes, are not present using this method.

Even though it may seem that approaches such as Linguistic Inquiry and Word Count (LIWC) [26], Affective Norms for English Words (ANEW) [27] or General Inquirer (GI) [28] are not as efficient for sentiment analysis, VADER's base sentiment lexicon was derived from those lexical features given they are already rated. The newly introduced lexicon was validated using a Wisdom-ofthe-crowd approach and then cleaned through a series of evaluations and validations to maintain the quality. Moreover, some generalizable heuristics were also created to incorporate wordorder sensitive relationships between terms. These heuristics include punctuation (!) to increase the magnitude of the sentiment intensity, capitalization to emphasize a sentiment-relevant word, degree modifiers (such as "extremely", "super"), contrastive conjunction "but" pointing a shift in sentiment polarity and examining the trigram preceding a sentiment-laden lexical feature. This allowed reaching the gold-standard list of lexical features used in VADER, which makes it the most convenient choice to classify the sentiment of tweets in the present dataset.

The classified tweets were then separated into different arrays depending on their sentiment and a word cloud was created using the WordCloud generator from Python to visualize words predominating in tweets belonging to each of the sentiments.

### 3.4. Exploratory Data Analysis

For the exploratory data analysis, the data of each of the vaccine datasets was manipulated using the Pandas library from Python. The time series for the number of tweets per day was developed using the time date Python module and the visualization was created using the matplotlib library. Further manipulation of data was carried out to prepare data for building the tree map of the percentage of tweets by state. The main results of these maps were coupled with the sentiment data to analyze sentiments on the days and states with the most tweets.

### 3.5. Social Network Analysis

To identify the users with the most influence in each of the vaccine datasets, the data from the user replies were used to build a network. The features of 'User ID' and 'In Reply to User' were used to build a table with the list of edges, with each edge referring to a connection of a user that replied to another user. Each of these users was referred to with their Twitter IDs and was set as the nodes of the graph. Since for this analysis we were interested in identifying the users with the most influence causing negative replies, the tweets that were selected for the graph were only those that had a negative classification. A single graph was built for every vaccine dataset using the Gephi software, filtering the nodes by their in-degree and showing the labels only for the nodes with in-degree ≥ 5.

## 4. RESULTS

### 4.1. Sentiment Analysis and Text Mining

Every tweet in each vaccine's dataset was assigned one of three different sentiments: neutral, positive, and negative. The number of tweets belonging to each class was normalized and is shown in Figure 2, where a general tendency towards neutrality can be observed. Figure 2 shows that Sinovac (54%) is the vaccine with the highest percentage of neutral tweets followed by Sputnik (48%), Pfizer (42%), CanSino (39%), and AstraZeneca (35%). Moreover, CanSino (39%) has a higher prevalence of positive tweets as opposed to Sinovac (28%), which had the lowest percentage. Sputnik and AstraZeneca presented the same percentage (32%), which was slightly lower than Pfizer (37%). Lastly, AstraZeneca (32%) had the highest percentage of negative tweets, followed by Pfizer (20%), CanSino (20%), Sputnik (19%), and Sinovac (17%). These percentages indicate an overview of people's reception towards a specific vaccine, but the context of why those sentiments arose is also a major topic to inquire about. Furthermore, it is important to mention that, although the following experiments were carried out using all the vaccine datasets, only the results for the AstraZeneca vaccine are shown in this paper. The summarized results for the remaining 4 vaccines can be visualized in Table 1 and Table 2.

The visualization of keywords on the created word clouds, along with existing evidence, allow the extraction of context that aid in the interpretation of the results from the sentiment analysis. For instance, in tweets related to positive sentiments found in the AstraZeneca dataset (Fig. 3), topics where people posted their absence of symptoms after they received their dosages could be found. Also, the millions of vaccines that would be shared by the USA with Mexico, with an initial batch of 1.5 million according to the Mexican Secretary of Foreign Affairs Marcelo Ebrard [19], was shown in the word cloud as well.

On the other hand, negatively classified tweets (Fig. 4) discussed the most commented side effect, which was the thrombosis that started appearing in several cases in European countries and of which news was spread quickly. Additionally, people appeared to be complaining about the government's decision not to suspend its use in Mexico. Most of these topics, including "blood clots", "side effects", "European countries", and "link between AstraZeneca", can also be observed in the 2-gram (Fig. 5) and 3-gram.

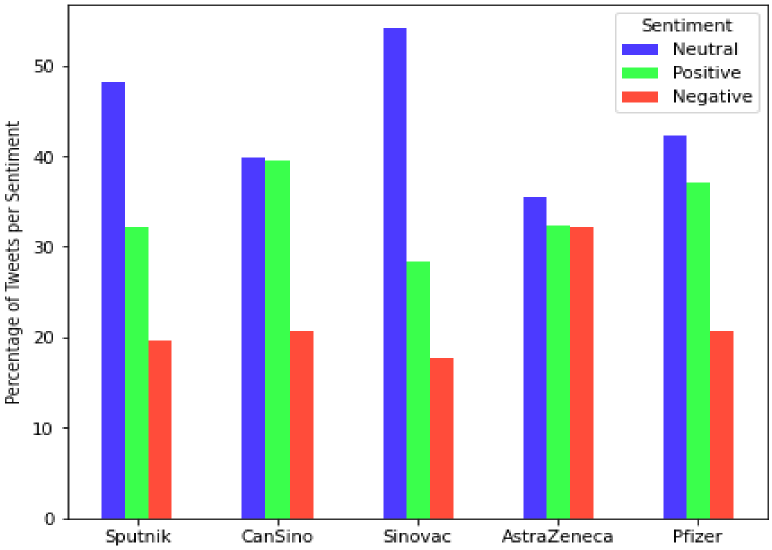


Figure 2. Obtained percentages of tweets classified by sentiment across five vaccines in Mexico.

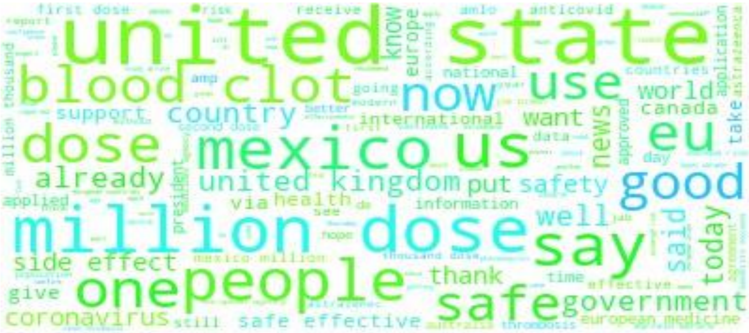


Figure 3. Word cloud based on positive AstraZeneca tweets.

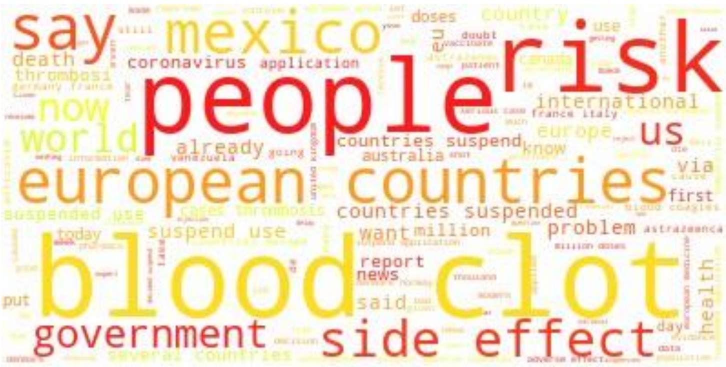


Figure 4. Word cloud based on negative AstraZeneca tweets.

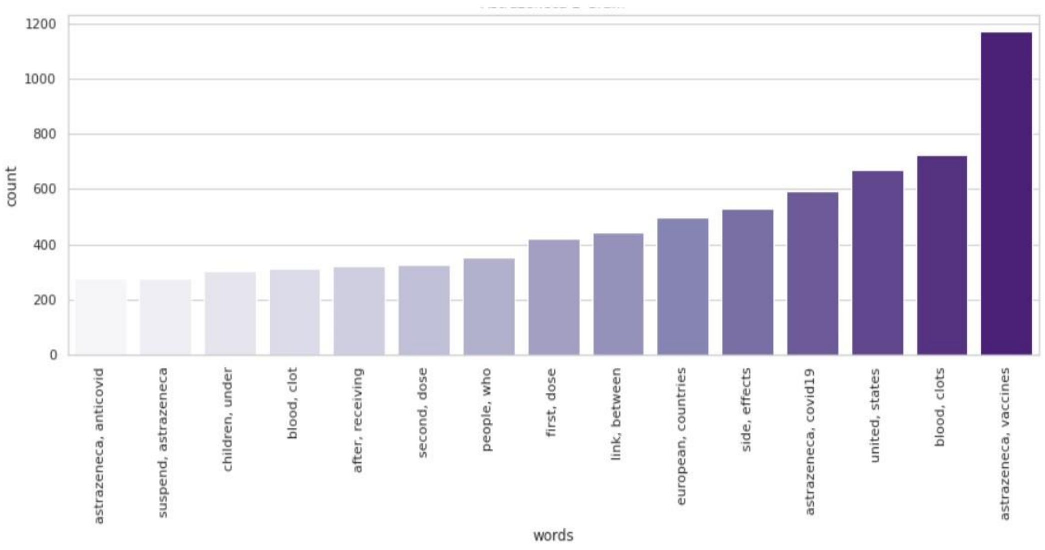


Figure 5. 2-gram on AstraZeneca vaccine dataset.

Similarly, the most important words appear in CanSino's positive word cloud, among which topics regarding its key feature of requiring only a single dose were found. Another topic involved people commenting about how their bodies reacted after receiving their unique dose. Being the most important news at that time was the packaging that would be carried out in Mexico at Queretaro's DrugMex Pharmaceutical Plant using CanSino's active substance [20]. Since this meant the vaccination progress would speed up, positivity in the tweets including this topic may be explained. On the other hand, tiredness was among one of the main symptoms people had days after their dose in connection to the negative sentiment tweets. In addition, negatively classified sentiments were linked to the vaccine's effectiveness, which was a very controversial issue when CanSino was first approved by the Mexican Federal Commission for the Protection Against Sanitary Risks (COFEPRIS). Likewise, those topics are detected in 2 and 3-Grams as "active substance", "single dose", "drugmex plant", "first three lots", and "over 60 years", where the latter refers to the decrease in the effectiveness in people over 60 years (Table 1).

In the case of the Russian vaccine Sputnik, negative sentiments also seemed to include several news. Among the most important news was the distribution of vaccines in the state of Campeche from apparently authorized retailers, in addition to the fact that, in the same state, vaccines were confiscated heading illegally to Honduras [21]. People also talked about their side effects after receiving the drug, but the previously mentioned news appeared to have more impact on the discussions. Similar to the previously discussed vaccines, the positively classified tweets from Sputnik, also showed people speaking up, either posting using their own Twitter profile or replying to other accounts, on the lack of symptoms or thanking governmental institutions for receiving their dose (Table 1).

Lastly, Sinovac's negative tweets referred to the users that were concerned over the vaccine's safety due to the death of an elderly person 40 minutes after he received his dose in the state of Guerrero. In other tweets, also about older adults, users were asking why Sinovac was still being applied to people over 60 years if only a few adults signed up to clinical trials, according to Europa Press news agency [22]. Besides this, another complaint was about the waste of vaccines due to the poor conditions in which they arrived in Mexico along with the bad temperature management that led to other vaccines being discarded as well. On the positive sentiment side, people talked about their experience days after receiving the vaccine and the millions of doses that were coming to Mexico [23] (Table 1).

Table 1. Sentiment analysis and text mining summarized results on CanSino, Pfizer, Sinovac, and Sputnik vaccine datasets.

| **Vaccine** | **WordCloud (Positive)** | **WordCloud (Negative)** | **2-Gram** | **3-Gram** |
|---|---|---|---|---|
| CanSino | single, dose, teacher, china, packaged, queretaro, ebrard, thank, health, advantage. | tired, teacher, emergency, cofepris, authorize, effectiveness. | single, dose; tired, vaccines; older, adults; first, batch; marcelo, ebrard; drugmex, plant; emergency, use; active, substance. | Over, 60, years; authorizes, emergency, use; first, three, lots; Coahuila, Chiapas, Nayarit. |
| Pfizer | us, first, vaccine, second, dose, well, thank, time, better, already, health, want, effectiveness. | shot, got, already, days, problems, doses, death, case, risk, side, effect. | Second, dose; first, dose; older, adults; third, dose; 60, years; astra, Zeneca; side, effects. | Over, 60, years; my, first, dose; south, african, variant; Mexico, new, lot; children, under, 12; young, people, between. |
| Sinovac | well, china, effective, thank, today, older, without, apply, reached. | effectiveness, poor, condition, tired, problem, hidalgo, temperature, adult, death, serious. | Second, dose; older, adults; first, dose; 60, years; after, receiving; pharmaceutics, vaccines, Pfizer; arrived, at. | Adults, over, 60; Mexico, has, received; minutes, after, receiving; has, an, effectiveness; 15, minutes, later. |
| Sputnik | dose, well, president, health, already, thank, according, effective, come, hope, case. | false, Campeche, report, Honduras, doses, via, case, problem, government. | Second, dose; first, dose; false, vaccines; not, know; would, have; private, aircraft; astra, Zeneca; new, lot; alberto, Fernandez; false, doses; direct, investment. | Direct, investment, fund; v, vaccines, confiscated; private, aircraft, at; has, an, effectiveness; san, pedro, sula; Campeche, international, airport; European, medicines, agency; 5000, false, doses. |

## 4.2. Exploratory Data Analysis

### 4.2.1. Number of Tweets per Day Time Series

In the Exploratory Data Analysis (EDA), the manipulation of data to obtain a time series of the count of tweets per day for each of the vaccines allowed us to identify trends within data. The number of tweets about the vaccines seemed to decrease on the weekend, resulting in a weekly trend of tweets having its peak in the middle of the week. However, a specific peak that corresponds to the day with the highest number of tweets per dataset could also be identified.

Most of the vaccines had the highest number of tweets in the middle of March (Table 2). For instance, the AstraZeneca dataset had the highest number of tweets on March 16 (Fig. 6) and the CanSino dataset on March 23. Also, the Sinovac and the Sputnik vaccine datasets both had their peak in the number of tweets on March 18, and, while the Pfizer dataset had a high number of tweets around these days, its peak was found on April 9.

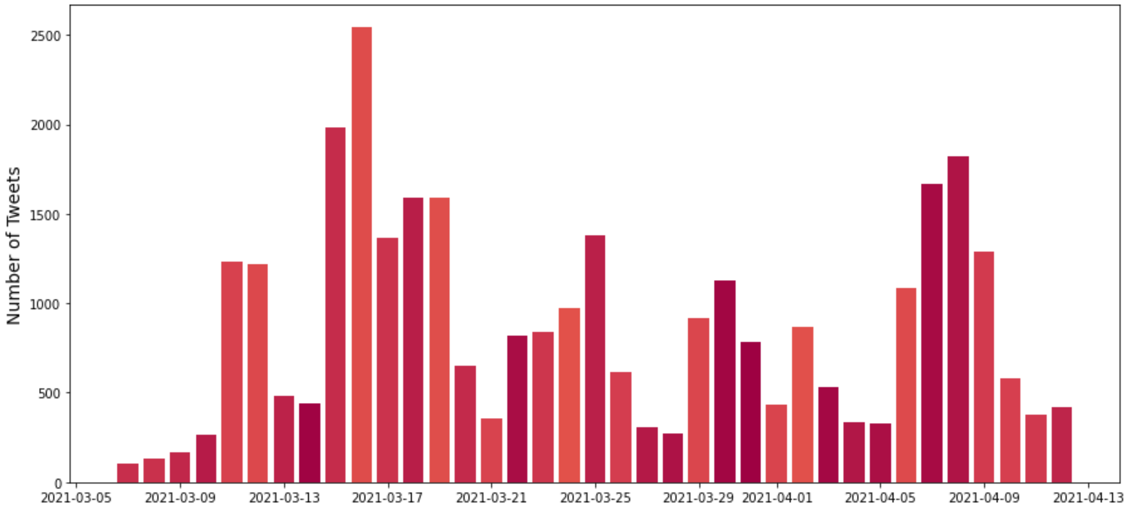


Figure 6. Number of AstraZeneca vaccine dataset tweets by day.

By tracing back on the events of the peak dates, it is possible to identify which type of events or news triggered a higher response on the Mexican population. Furthermore, by identifying the sentiments among the tweets on these peak dates, the nature of the response can also be elucidated. For instance, the AstraZeneca vaccine had a predominance of negativity on the day with the most tweets (Fig. 7). This was the same case for the CanSino vaccine, although the difference in the percentage of negative tweets was higher in the CanSino dataset (45%) than in the AstraZeneca dataset (38.5%). The remaining three vaccines had predominance for neutrality, resulting in percentages of 70.71, 56.56, and 38.88, for Sinovac, Sputnik, and Pfizer, respectively. In these cases, the percentage of neutral tweets was significantly higher in the Sinovac and the Sputnik vaccine datasets than in the Pfizer dataset, and, also, the second most predominant sentiment was negative in Sputnik and Sinovac datasets as opposed to the latter, in which there was a higher percentage of positive tweets (Table 2). This provides evidence that the immediate response to events about news for each of the vaccines had a positive nature for the Pfizer vaccine and a negative one for the rest of the vaccines, specially CanSino and AstraZeneca.

### 4.2.2. Percentage of Tweets by State

Data was further manipulated to view how the percentage of tweets obtained for each of the vaccine datasets was divided into each of the Mexican states. The results were quite similar among the five datasets (Table 2). The states of Puebla, Querétaro, and San Luis Potosí had the highest percentage of tweets in the AstraZeneca (Fig. 8) and the Sputnik datasets. Similarly, CanSino and Sinovac also had among their top 3 the states of Puebla and Querétaro, differing only by the states of Guanajuato and Oaxaca, respectively. Finally, the Pfizer vaccine dataset also had Puebla and Oaxaca among its top 3 states, along with Veracruz. These results indicate that the states that seemed to have the most activity on Twitter regarding the vaccines were Puebla and Querétaro.

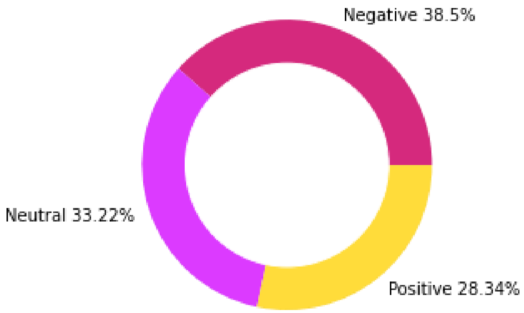


Figure 7. Percentage of tweets per sentiment classification on AstraZeneca vaccine dataset on March 16.

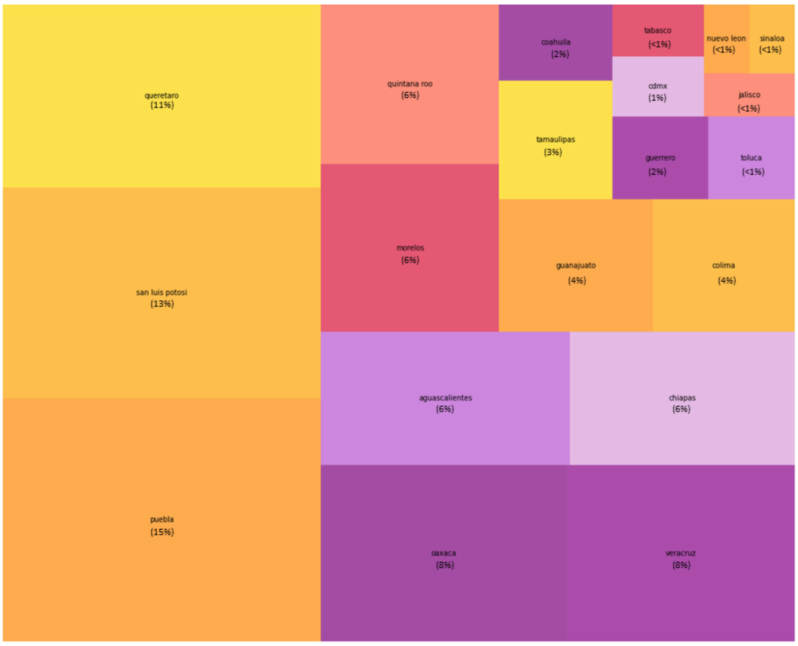


Figure 8. Tree map of percentages of tweets per state on AstraZeneca vaccine dataset.

Furthermore, the sentiments obtained were also coupled with these results (Table 2). For instance, the predominant sentiment on the top 3 states with a higher percentage of tweets was negative for the AstraZeneca vaccine dataset (Fig. 9). This was the same case for the state of Guanajuato on the CanSino dataset, which was also among the top three states with the highest percentage of tweets for this vaccine. However, the rest of the vaccine datasets and their respective top 3 states had a high predominance for neutral sentiments; and for all the vaccines except AstraZeneca and CanSino, the second most predominant sentiment was positive. These results may indicate that the overall response of the most active states regarding the vaccines was negative for AstraZeneca and CanSino and positive for Sputnik, Pfizer, and Sinovac.

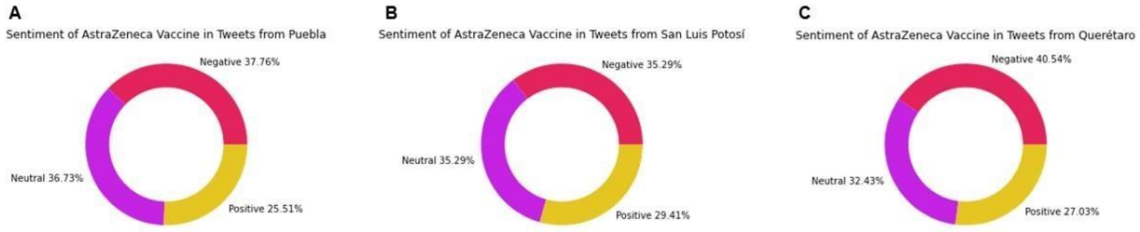


Figure 9. Percentage of tweets per sentiment classification on top three states with the highest tweet count on AstraZeneca vaccine dataset.

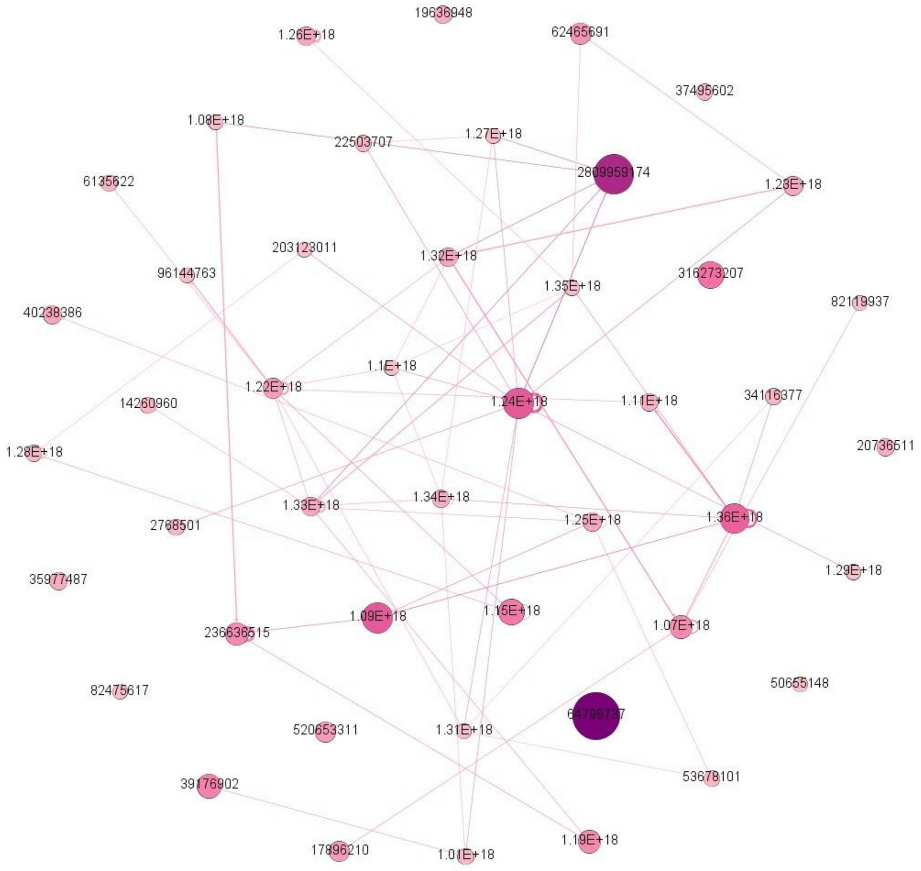


Figure 10. Social Network Analysis of user replies in AstraZeneca vaccine dataset. Graph is filtered by in-degree ≥ 5, in which labelled nodes correspond to this equality. Numbers in nodes represent the ID of the Twitter users: 64798737 is Marcelo Ebrard, 316273207 is Dr. Alejandro Macias, and 236636515 is Joaquin López-Dóriga.

## 4.3. Social Network Analysis

The Social Network Analysis was based on the user replies from each of the vaccine datasets. A network was generated per vaccine and was conformed of tweets that were classified as negative. Filtering the network allowed the identification of the users that had the most negative replies about the vaccines. For example, the Mexican Secretary of Foreign Affairs, Marcelo Ebrard, was one of the users with the most replies in all the vaccine datasets. Similarly, Dr. Alejandro Macías, a researcher from the University of Guanajuato, appeared as one of the nodes with the most replies in the AstraZeneca (Fig. 10), Sinovac, Sputnik, and CanSino vaccines. A famous Mexican journalist called Joaquín López-Dóriga was also highly present in the negative replies from the CanSino and AstraZeneca vaccine datasets.

Other figures such as researchers and politicians were also quite present in the vaccine datasets. For instance, Lilly Téllez, a Mexican journalist and politician had several replies in the Sputnik dataset. Also, the Quintana Roo Governor Carlos Joaquín appeared as a large node in the Pfizer dataset, while Dr. Alma Maldonado, a researcher, and professor from the University of Mexico was highly present in the CanSino dataset. The results from these networks indicate that politicians and researchers are highly involved in negative responses regarding each of the available vaccines in Mexico. Further investigations on the nature of the tweets that were highly replied to may provide an insight into the type of allegations that have the greatest impact on the target users.

Table 2. EDA and Social Network Analysis summarized results for 5 vaccine datasets.

| | **Sinovac** | **Pfizer** | **CanSino** | **Sputnik** | **AstraZeneca** |
|---|---|---|---|---|---|
| **Mean (tweets/day)** | 166 | 605 | 93 | 237 | 856 |
| **Max tweet count** | 478 | 1273 | 622 | 838 | 2507 |
| **Day with highest tweet count** | March 18, 2021 | April 9, 2021 | March 23, 2021 | March 18, 2021 | March 16, 2021 |
| **State with highest tweet count** | Puebla | Puebla | Puebla | Queretaro | Puebla |
| **Sentiment (highest tweet count day)** | Neutral | Neutral | Negative | Neutral | Negative |
| **Sentiment (highest tweet count state)** | Neutral | Neutral | Neutral | Neutral | Negative |

| | | | | | |
|---|---|---|---|---|---|
| **Main subjects in social network** | Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher) | Marcelo Ebrard (Politician); Carlos Joaquín (Politician) | Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Dr. Alma Maldonado (Researcher); Joaquin Lopez-Doriga (Journalist) | Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Lilly Téllez (Politician/Journalist) | Marcelo Ebrard (Politician); Dr. Alejandro Macías (Researcher); Joaquin LopezDoriga (Journalist) |

## 5. DISCUSSION

The analysis carried out in this research demonstrates an overview of people's reception towards a specific vaccine. The classes assigned to each tweet in every dataset allowed the visualization and analysis of the different types of reactions that were encountered for each of the vaccines, and the reason of those sentiments could be modelled using word clouds that enabled the identification of keywords that could provide a clue on the topic that was being discussed. For instance, side effects appear frequently among negative sentiment keywords, as well as complaints against the government, or reactions to the most shocking news affecting the vaccines, not only in Mexico but also in other parts of the world. On the other hand, positive tweets involved appreciation to the government for supplying the vaccines and spoke positively about the safety and effectiveness of the different vaccines.

To provide more context to the keywords found on the word clouds, 2-grams, and 3-grams for each of the vaccine datasets were also created. These n-grams allowed us to visualize the most important phrases from which a context can be extracted without the need for modelling using a Topic Model such as non-negative matrix factorization (NMF) or Latent Dirichlet Allocation (LDA). However, LDA was performed as a validation method for the relationship found between the topics on the word clouds and the n-grams. Given their similarity, LDA results are not shown in this paper. Furthermore, since the results of the n-grams were not divided into positive and negative tweets as in the word clouds, the small phrases that appeared depended on how positively or negatively people talked about the specific vaccine. For instance, since AstraZeneca had more negative tweets, the most recurrent n-grams were related to the blood clots side effect, while the case of CanSino, which was the vaccine with the most positive tweets had good comments regarding its single-dose feature. The identification of these sentiments and topics is relevant as social networks have a tremendous impact on mass behavior and tendencies. Consequently, although positive comments could have the effect of promoting vaccination, negative comments could also affect the decision of the population not to apply their vaccine.

A brief EDA was carried out to have a general visualization of how the tweets from each vaccine dataset were distributed. First, with a segmentation of the count of tweets per day, we were able to obtain the days with the most tweets of any sentiment for each of the vaccine datasets. By tracking back on the news or important announcements, the tweets from our database, and the results from the previous analyses, we could identify several situations in which the users reacted and tweeted the most. For instance, in the case of the Sinovac vaccine, the day with the most tweets,was the day where Mexico received a batch of a million doses [23], for which people had more positive than negative reactions. This was also the case for the Pfizer vaccine, where the main announcements involved the arrival of a new batch of vaccines, and the affirmation of the President that with this new arrival, the doses for all Mexican older adults would be sufficed

[24]; the tweets on this day were also mostly positive. Moreover, in the case of CanSino, announcements that were made on the day with the most tweets were related to the liberation and immediate application of the vaccines in rural regions and the acquisition of more CanSino doses that were packaged in the state of Querétaro [25]. As opposed to Sinovac and Pfizer, these announcements had a higher percentage of negative responses. Similarly, on the day with the most tweets on the AstraZeneca vaccine, the main announcement was the negotiation of Mexico with the US to obtain these vaccines [19], which also had a mostly negative response. Finally, the day with the most tweets in the Sputnik dataset matched the day of the announcement of the discovery of fake vaccines in the state of Campeche [21]. Although most of these tweets had a neutral tone, the negative responses predominated over positive ones. This information leads to the conclusion that overall, depending on the previous "reputation" of the vaccine, announcements involving the arrival of more doses would define the polarity of users' reactions towards it.

Also, the datasets were used to explore the states that tweeted the most about each of the vaccines. Overall, it was found that the number of tweets was concentrated in the southeast and north-central regions of Mexico. While most results indicated neutrality and positivity towards the vaccines, there was a notorious exception in the case of the AstraZeneca and CanSino vaccines, whose tweets from their top three states with the highest number of tweets had a negative-neutral tendency. While these results provide evidence on the regions of the country that are the most active in Twitter regarding the vaccines, it would be necessary to further study the nature of the tweets by each of the states to define the impact that these states are having on the rest of the country. Furthermore, it would be necessary to validate whether the datasets are balanced in terms of the number of tweets that were recovered by state using the API.

The results from the Social Network Analysis let us identify those politicians, journalists, and researchers that were the most influential figures on the vaccine datasets. For instance, the Mexican Secretary of Foreign Affairs, Marcelo Ebrard, was the node with the highest in-degree in almost all the datasets. This was expected as the Secretary gives several announcements on the negotiations and transportations of the different vaccines. A similar effect was visualized in the case of the famous Mexican journalist Joaquin López-Dóriga. However, politicians Lilly Tellez and Carlos Joaquín, as well as researchers Alejandro Macias and Alma Maldonado, could represent specific targets within the datasets whose comments trigger a response on the community. Since the networks were built based on the negatively classified tweets, we can assume that these people triggered a negative response, however, deeper analysis on each of the subjects to define what is the exact nature of their impact and which characteristics they present to create this impact in the community. It is important to mention that, although targets such as these could be extracted from the networks, noisy users were also found. For instance, in the case of the Sputnik vaccine, one of the nodes with the highest in-degree was that of an influencer named "Sputnik". These types of cases introduce noise to the analysis and should therefore be considered in future work.

Overall, the sentiment analysis coupled with the text mining, the EDA, and the social network, allowed us to have a deep overview of each of the datasets created for each of the vaccines available in Mexico. This approach resulted useful to gain insight into the reactions and behaviors of the Mexican population towards the vaccines and to open the landscape to future investigations where misinformation sources can be predicted, and to further implement measures that may slow down the spread of fake news or allegations that may be harmful to the community. Actions that can be taken to further improve the current method include refining the geolocation of the queries from Twitter, curating users that may include noise into the datasets, including a wider variety of sentiments for classification, and creating social networks that demonstrate the polarity of the community. Further work on vaccine follow-ups is also encouraged.

## 6. CONCLUSION

The method employed in this work for evaluating the sentiments of Mexicans towards the currently available COVID-19 vaccines was effective in gaining an insight into the behavior and opinions of the Mexican population. Using the VADER toolkit allowed the classification of tweets into three different sentiments, providing a format with which the data could be further analyzed. This method combined with text mining, exploratory data analysis, and social network analysis allowed the identification of key words, dates, geographical regions, and people that were highly represented by the data. While improvements such as expanding the number of sentiments from the tweets and refining translation can be made, the current analysis provides a framework that contributes to the understanding of population behavior and to the identification and avoidance of possible misinformation.

## REFERENCES

[1] Daniel Allington, Bobby Duffy, Simon Wessely, Nayana Dhavan, and James Rubin. Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. Psychological Medicine, page 1–7, 2020.

[2] World Health Organization. Ten threats to global health in 2019. https://www.who.int/newsroom/spotlight/ten-threats-to-globalhealth-in-2019, Mar 2019.

[3] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. PloS one, 5:e14118, 11 2010.

[4] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarization. In Proceedings of the 24th International Conference on World Wide Web, WWW ’15 Companion, page 355–356, New York, NY, USA, 2015. Association for Computing Machinery.

[5] J. Leung, J. Y. C. Chung, C. Tisdale, V. Chiu, C. C. W. Lim, and G. Chan. Anxiety and panic buying behaviour during covid-19 pandemic-a qualitative analysis of toilet paper hoarding contents on twitter. International Journal of Environmental Research and Public Health, 18:1–16, 2021.

[6] J. Talbot, V. Charron, and A.T. Konkle. Feeling the void: Lack of support for isolation and sleep difficulties in pregnant women during the covid19 pandemic revealed by twitter data analysis. International Journal of Environmental Research and Public Health, 18:1–12, 2021.

[7] A. Jabeen, S. Afzal, M. Maqsood, I. Mehmood, S. Yasmin, M.T. Niaz, and Y. Nam. Anlstm based forecasting for major stock sectors using covid sentiment. Computers, Materials and Continua, 1, 2021.

[8] Abdur Rasool, Ran Tao, AlimarjanKamyab, and Tayyab Naveed. Twitter sentiment analysis: A case study for apparel brands. volume 1176, page 022015, 03 2019.

[9] Sarah Shukri, Rawan Yaghi, Ibrahim Aljarah, and Hamad Alsawalqah. Twitter sentiment analysis: A case study in the automotive industry. 11 2015.

[10] Wu He, ShenghuaZha, and Ling li. Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33:464–472, 06 2013.

[11] Muna Al-Razgan, Asma Alrowily, Rawan L. Al-Matham, Khulood M. Alghambdi, MahaShaabi, and Lama Alssum. Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by saudi women. Technology in Society, 65, 05 2021.

[12] E. Rajesh Kumar and K.V.S.N. Rama Rao. Sentiment analysis using social and topic context for suicide prediction. International Journal of Advanced Computer Science and Applications(IJACSA), 12, 2021.

[13] Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies, 2018.

[14] Kaushal Kumar Bhagat, Sanjaya Mishra, Alakh Dixit, and Chun-Yen Chang. ”public opinions about online learning during covid-19: A sentiment analysis approach. Sustainability, MDPI, 13:1–12, 03 2021.

[15] Katharina Toeppe, Hui Shenghua Yan, and Samuel Kai Wah Chu. Repurposing sentiment analysis for social research scopes: An inquiry into emotion expression within affective publics on twitter during the covid-19 emergency. Diversity,Divergence,Dialogue, page 396–410, 02 2021.

[16] Pristiyono, MulkanRitonga, Muhammad Ali Al Ihsan, AgusAnjar, and Fauziah Hanum Rambe. Sentiment analysis of covid-19 vaccine in indonesia using naıve bayes algorithm. IOP Conference Series: Materials Science and Engineering, 1088, 2021.

[17] Sameh N. Saleh, Samuel A. McDonald, Mujeeb A. Basit, Sanat Kumar, Reuben J. Arasaratnam, Trish M. Perl, Christoph U. Lehmann, and Richard J. Medford. Public perception of covid-19 vaccines through analysis of twitter content and users, 04 2021.

[18] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.

[19] ’exitosa’ la negociacion con eu para que mexico obtenga vacunas de astrazeneca: Lopez-gatell, Mar 2021.

[20] Marittza Navarro. Drugmex, la empresa en queretaro que envasara la vacuna china contra el covid19, 02 2021.

[21] ElıasCamjahi and Marıa R. Sahuquillo. Decomisadas en mexicomas de 5.000 dosis falsas de la vacuna rusa sputnik v.

[22] MSN Noticias. La omsaprobo el uso de emergencia de la vacuna de sinovac, 06 2021.

[23] Cesar Arellano Garcıa. Llega a mexico un millon de vacunas de sinovac, Mar 2021.

[24] Mexico cuenta con todas las dosis necesarias para terminar de vacunar a personas adultas mayores el 20 de abril: presidente, Apr 2021.

[25] Vacunas cansino se usaran inmediatamente en los 32 estados: Lopezgatell, Mar 2021.

[26] Pennebaker, James & Francis, Martha & Booth, Roger. Linguistic inquiry and word count (LIWC). 1999.

[27] Bradley, M. and P. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. 1999.

[28] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates.The General Inquirer: A Computer Approach to Content Analysis.MIT Press, 1966.

## AUTHORS

**Jessica Salinas** is a Biotechnology Engineer from Monterrey, Mexico, currently pursuing a master’s degree in Computer Science at Tecnológico de Monterrey, Mexico. Her main work experience has been working in research laboratories and IT consulting at Accenture Technology, Mexico. Nevertheless, Jessica also has experience in teaching languages and in developing IT workshops for university students. Her current interests rely mainly in Bioinformatics applied to health and plant science, Data Science, Data Analytics, and Data Visualization.

**Carlos Flores Munguia** received the B.S. degree in Computer Systems engineering from Tecnologico Nacional de Mexico, Tamaulipas in 2019, graduating with honours thanks to achievements obtained during his university career. Carlos is currently earning a master’s degree in Computer Science at Tecnologico de Monterrey. He has served as a teacher of different workshops with the purpose of helping students to develop talents not explored by the institution. He has also participated and won in innovation competitions, for which he has represented his city at state level. His research interests include the success in the application of genetic algorithms in real world problems, high performance computing, and Few-Shot Learning as an alternative to data-intensive Deep Learning approaches.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN ANXIETY AND STRESS REDUCING PLATFORM BASED ON MINIGAMES AND EMOTIONAL RELEASE USING MACHINE LEARNING AND BIG DATA ANALYSIS

Selina Gong[1], John Morris[2] and Yu Sun[2]

[1]University High School
4771 Campus Dr, Irvine, CA 92612
[2]California State Polytechnic University, Pomona, CA, 91768

### ABSTRACT

*Today's students are faced with stress and anxiety as a result of school or work life and have added pressure from social media and technology. Stress is heavily related to many symptoms of depression such as irritability or difficulty with concentration as well as symptoms of anxiety like restlessness or feeling tired. Some of these students are able to find a healthy outlet for stress, however other students may not be able to. We have created a program where students will be able to destress and explore their emotions with the help of suggestions from our system based on previously explored thoughts. Our program uses machine learning to help students get the most effective stress relief by suggesting different mental health exercises to try based on input given by the user and provides emotional comfort based on the user's preferences.*

### KEYWORDS

*relaxing, destress, game, journal.*

## 1. INTRODUCTION

The two biggest factors of teen depression are stress and technology [1]. Over 4 million children aged 3-17 in the US have either anxiety or depression, and this number has only gone up in recent times. Many teenagers spend over 7 hours on screen (from phones to computers and tvs) not including the amount of time they spend on schoolwork. In these hours, students frequent entertainment sites such as social media and video watching sites, which may have negative impacts on the student's self-esteem [2]. The Covid-19 pandemic also placed pressure on teenagers due to the sudden spike in fear and isolation. Not only that, but when school resumed, many students found themselves facing a higher level of academic stress from having to adjust to online learning as well as the inefficiency of teaching through online meetings. Having to self quarantine for a long period of time also ruins many teenagers' routines, specifically their sleep schedules, which could lead to an imbalance in hormones. All of these factors put teenagers at risk for depression, which is characterized by a loss of energy, poor school performance, self-harm, and even suicidal thoughts [3]. The issue of mental health has stigma surrounding it, making it difficult for teenagers to discuss the topic with people in their immediate lives. Instead, it is easier for students to go online, where no one who truly knows who they are and vent about whatever happened in their lives that day.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 179-192, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111416

There are self-care applications on the app store, as well as relaxing games, but our program aims to combine them both and help the user relax and de-stress while also helping the user explore emotions. Some applications that aim to help the user self-reflect as a means of relieving stress rely entirely on the user being willing to continue putting in the effort of thinking about their own emotions. However, this assumes that the user has enough energy to open themselves up and discuss the most vulnerable parts of their emotions. And this is also assuming that the user will remember to use the application in the first place. Not only that, but some of these applications are simply online journals that ask the user questions and store their answers [4]. This kind of application is marketed toward helping depressed or stressed users, but does not provide any incentive for the user to continue since they are so reliant on user input and the users it is marketed toward do not have the energy to continue a long journey of self-reflection without instant gratitude. A second problem would be that the user input is not used to help the user. Users are allowed to customize their own journal, but that's about it. This results in users having to put in their own effort to analyze their own responses and figure out the best way to ease their stress. Once again, considering that the target audience is students with high levels of stress and depression, not all users will have the time to do so.

In this paper, we research if our game is useful in helping students destress. We believe that this method may not be as successful with helping the user understand themselves better in comparison to other applications that are more focused on guiding the user through exploring their own thoughts, however our method aims to help the user either relax or get happier through a few fun minigames. Our goal is to optimize relaxation for users through the collection of data that they input into the game, specifically the journal minigame. Some features included in our method are a journal that records user input and multiple minigames that do not have a time limit pressure or a large pressure for failure. We also utilize the tendency for younger adults to relate with inanimate objects in our method with our number of Non-Playable Characters (NPCs) [5]. Since the recent Covid-19 pandemic required most citizens to self-quarantine, there were increased feelings of loneliness throughout the world. Having a digital "friend" gives the users a character to relate to and through this, they may be able to grow or come to a new understanding of the self. The NPCs in our game each have their own personalities and elaborate backstories that users can read about. Finishing certain tasks may also trigger a message from the NPCS. Other methods may not have an incentive for success and may not give the user a sense of attachment. We also wanted to help users relax through lowering breathing and heart rates [6]. Our game uses low BPM music in every scene to achieve this.

We have developed a program that analyzes whether a user will have had an improvement in mood or not depending on the user's frequency in playing a game, the amount of time spent playing that game, and the user's score. We show the usefulness of our approach with an experiment that logs many aspects of user actions.

The rest of the paper will follow this outline: Section 2 describes challenges we encountered during the process of creating and carrying out this experiment, Section 3 presents related works, Section 4 explains the methodology we used and the solution we came up with, Section 5 evaluates the experiment and provides extra details, and Section 6 gives concluding remarks.

## 2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

### 2.1. Making the Game Interesting

Since we were making a game, we had to consider how we would gain the user's attention and interest for an extended period of time. We didn't want to make this a method that users would use a few times and forget about. We wanted to give the users a sense of accomplishment that is relatively easy to get, but still requires a bit of effort to obtain. With this, we came up with the scoring system and friendship points. We created NPCs that each have their own unique personality and backstory so users will want to become better acquainted with them [7]. However, we did not want to make it so that the user would just click through the stories without much gratitude. We decided that these backstories would be locked behind a friendship point cost. Friendship points would be easily obtainable through daily use of our game. In the end, we laid out a plan for how to retain user interest in our game.

### 2.2. Scoping an Accurate Data-Driven Model

We didn't want to have to continually ask our users for their emotions before and after their experience with our game, because we would not always be there to ask and users were not guaranteed to always remember to complete the journal that logs emotion at the correct times. In order to be able to determine whether the user was able to relieve stress by playing our game without having to ask them to log it, we wanted to create a data-driven model that was relatively accurate. We had to test multiple machine learning algorithms each with their own parameters and specialized abilities. We wanted to determine which machine learning algorithms would prove to be more successful in determining whether the user experienced stress relief [8]. Since there are many machine learning algorithms available for us to use, we had to research which ones would fit best for our problem. Not only that, but most algorithms we chose had multiple parameters. In order to get the best accuracy, we needed to change each parameter and test them individually to determine which factors together would return the best accuracy while still being reasonable.

### 2.3. Figuring Out how to Make Matched Objects

One minigame, the matching game, contained a function that required the game to open a different scene, the journal scene, and then return to the minigame scene. The problem was that whenever users returned to the matching scene, everything got reset since everything that users could interact with was a clone, so all of the NPCs left the screen and came back in and any previously matched items reappeared. However, we needed to make it so that they only did not reset if the user was returning from the journal scene. For this, we needed to ensure that the game kept track of which scene the user was returning from and we needed the scene to hold references to which NPCs were in the scene previously [9]. In the end, we created a system that kept track of which NPCs were previously in the scene, the position they were in, and which items were already matched.

## 3. RELATED WORK

*Identification and categorization of digital game experiences: a qualitative study integrating theoretical insights and player perspectives* [10] aims to analyze the effects of different kinds of games and the experience its players get from them. This work summarized that some of the main reasons why someone would play a game are: having free time, wanting to relax, or wanting to socialize. Some of the experiences the participants had related to sound, exploration, and relaxation. These findings were very broad in terms of finding the experiences of different gamers when playing, and were not as specifically focused on games and their effects on relaxation.

*Stress Relieving Video Games: Creating a Game for the Purpose of Stress Relief and Analyzing Its Effectiveness* [11] successfully analyzes the relaxing effect of two different kinds of games and de stressing exercises on a variety of people. One of the games was a researcher made game and the other was Tetris. The researcher found that all three methods of relaxation mostly resulted in a decrease in heart rate, save for an outlier, however the participants themselves were split on whether they felt distressed after the researcher made a game and the breathing exercises. The participants who played Tetris did feel distressed, though. This researcher did something very similar to what we did, however they included different kinds of games with a very few participants [12]. Having only six participants may have led to skewed results.

*Implementation of serious game techniques in raising the social awareness of the depression disease* [9] analyzes how effective a serious depression awareness game was in informing users of the risks and effects of depression. The study had a large participant group, 89 participants, that showed that the majority of the participants were uninformed about depression before the game and most participants felt much more informed about depression after the game.

## 4. SOLUTION

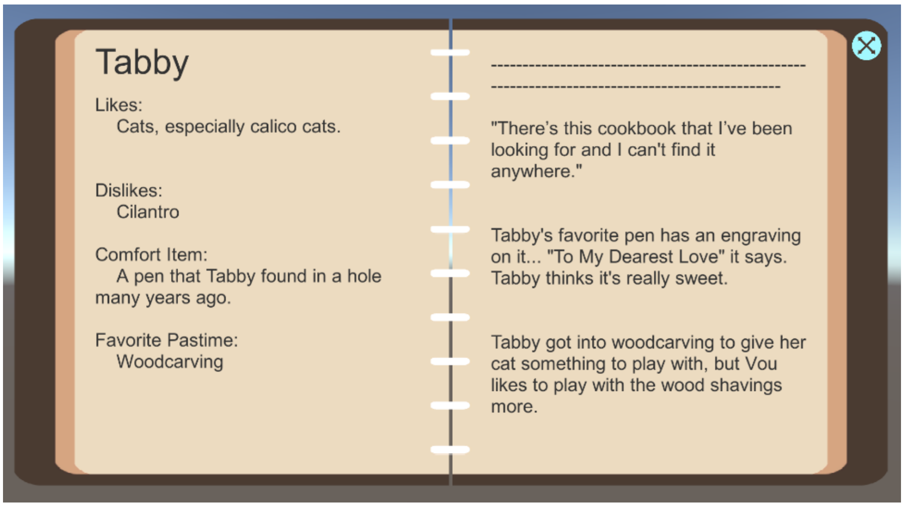
Flory is a video game meant to provide relaxation and happiness through the use of three different minigames and a small low-poly world to explore, this is also an adaptive and interactive gaming platform for depression and stress relief using machine learning.

There are many Non Play Characters (NPCs) in this game, each with unique backstories and personalities [11]. The first minigame is a journal minigame meant to help the user explore or release any pent up emotions.

The second minigame is a click and drag matching game. The user will help NPCs find items they want to buy from a store's shelves.

The third minigame is a popping game. Flower seeds will pop up on screen and when tapped, the flower will bloom and fall to the ground. The goal is to pop enough flowers to fill the field.

The game uses machine learning to recognize which NPCs the user likes to spend friendship points on and will prioritize use of those NPCs. Through these three minigames, the game will learn to process player emotions using machine learning basics with data collection and data training. Eventually, the game will be able to provide recommendations for which minigame to play based on the emotion input in the journal minigame. The ultimate goal is to use machine learning as the backend for emotional & mental support through a gaming experience.



Tabby

Likes:
Cats, especially calico cats.

Dislikes:
Cilantro

Comfort Item:
A pen that Tabby found in a hole many years ago.

Favorite Pastime:
Woodcarving

------------------------------------------------
---------------------------------------------

"There's this cookbook that I've been looking for and I can't find it anywhere."

Tabby's favorite pen has an engraving on it... "To My Dearest Love" it says. Tabby thinks it's really sweet.

Tabby got into woodcarving to give her cat something to play with, but Vou likes to play with the wood shavings more.

Figure 1. Mini game 1

In minigame 1, the journal minigame, users are asked a set of questions related to their day and emotions that they feel. This will help the users vent any unhealthy emotions in a way that doesn't harm anyone around them and will also prompt users to reflect upon themselves and grow as a person. The user's answers are recorded and the user will be allowed to go back to previous days to recall what they were feeling before. The user is allowed to complete one entry per day and each complete entry awards the user with 10 friendship points. (See Appendix A for code segment.)

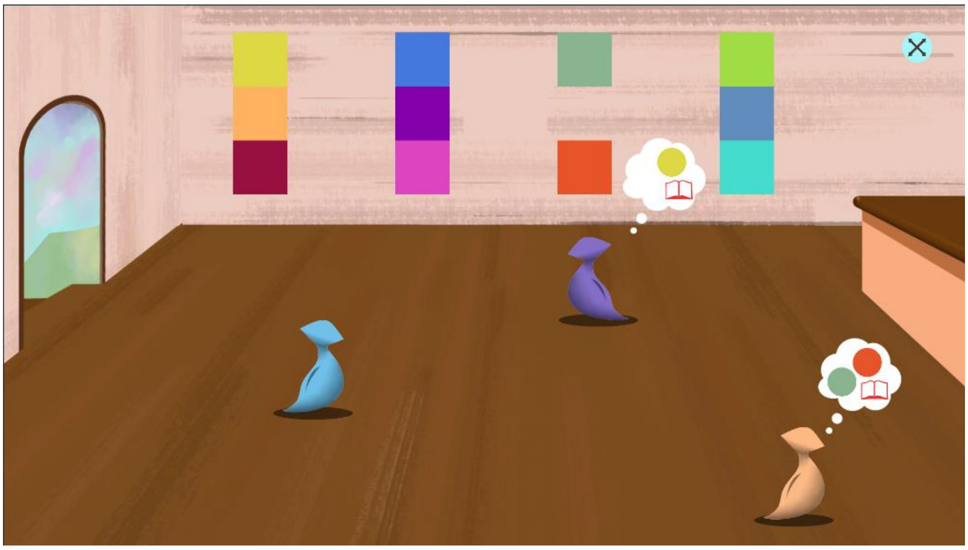


Figure 2. Mini game 2 (1)

The game also includes a matching minigame featuring the NPCs and different items that they want to buy. Each NPC has their own personal journal where the user can learn the NPCs likes and dislikes, comfort items, and favorite pastime(s) [12]. The second page of the NPCs journal is full of entries that can be unlocked for 15 friendship points that contain more insight into the NPCs personality, but more importantly, these entries will tell the user what items that specific NPC wants to buy. (See Appendix B for code segment.)

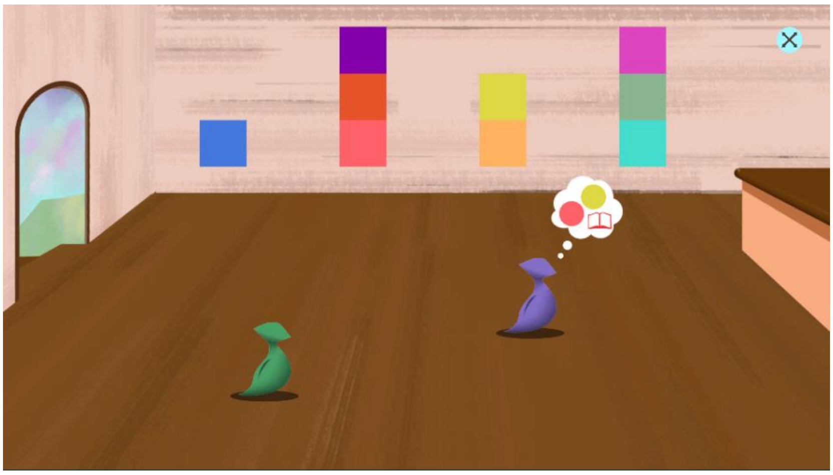


Figure 3. Mini game 2(2)

Since these entries need to be unlocked, it is entirely possible that an NPC featured in the matching game will want to buy an item that the user does not know. In the minigame, items will appear on shelves. Three random NPCs will come into the store and ask for two items in a thought bubble. One item will be a random general item, and one will be an NPC unlockable item. If the item is locked, then the user will not see it. When an item is dragged off of the shelf and near the correct NPC, the item and its icon in the NPCs thought bubble will disappear. Once both items have been matched, the NPC will leave the room.

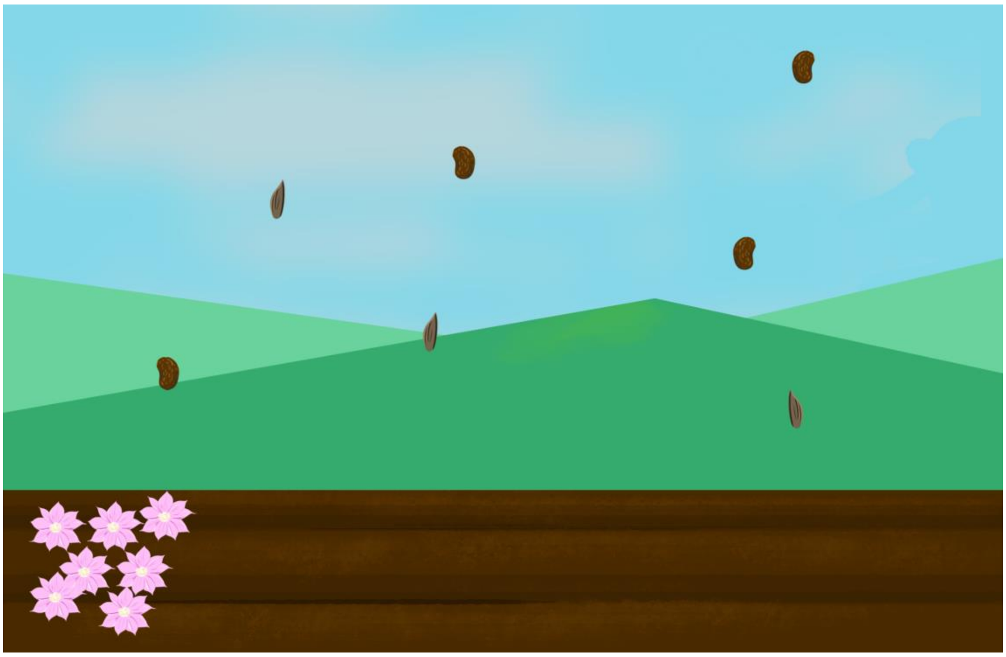


Figure 4. A screenshot of mini game 3

In the flower popping minigame, seeds will randomly spawn on screen. There will be a maximum of 10 seeds on screen at any time and a minimum of 5. Clicking on a seed will cause that seed to bloom into a flower and fall into the flower field at the bottom of the screen. When the user pops enough flowers to fill the flower field, the user wins and is given the option to continue if they would like. Each 20 flowers popped will give the user 10 friendship points.

## 5. EXPERIMENTS

Five high school students participated in this study. Participants ranged from 14 years old to 18 years old. All participants said they felt stressed from school work and extracurriculars. Three participants' main source of stress came from college applications.

### 5.1. Experiment 1: Does our game reduce student stress?

To evaluate the effectiveness of the game in improving the user's emotion, we asked 5 students to play the game, logging their emotions before and after using the journal minigame. In other words, they would be playing the game as intended and all data would be collected as they were playing. Students were encouraged to ignore time limits and the score was not displayed in order for students to not be pressured by the game itself. They were also not watched while playing the game. In this way, users were able to fully immerse themselves in the game and relax.

There are four pieces of data that are logged and the fifth piece will be determined from the other pieces of data. In the journal game, users are able to log their current emotion twice per day. Participants were instructed to fill out the journal once at the beginning of their experience and once at the end. This is to determine whether the participant had an improvement in mood and/or stress relief.

In the matching minigame, data is automatically logged at the end of the game. The timestamp at which the user completed the game, the duration of the game and the score as well as whether the user played a full game are logged. The score of the game as well as whether the game was completed is also logged. A perfect score is 12 points.

Table 1 below is the data for an example participant. The emotion before they started playing the game was sad. They played the matching game twice with a combined time of 62 seconds and a combined score of 16. They did not complete the second game, though.

Table 1. Example of Data Collected In-Game

| Scene Name | Time | Log Type | Extra |
|---|---|---|---|
| Journal | 1629320087.88526, | Emotion Before: sad | |
| Matching Game | 1629320110.57972, | Duration: 39 | Score: 12, True |
| Matching Game | 1629320127.3878, | Duration: 23 | Score: 4, False |
| Journal | 1629320155.85095, | Emotion After: neutral | |

Table 2 below shows the data of all eight students that participated in this study. Of the eight students, five of them had an improvement in mood after playing the game. The average time of the five participants who had an improved mood was 28.9 seconds per round with an average score of 10.9 points per round.

Table 2. User-study data from participants

| Participant # | Frequency | Total Duration (s) | Total Score | Mood Improved? |
|---|---|---|---|---|
| 1 | 1 | 15 | 12 | No |
| 2 | 3 | 63 | 34 | Yes |
| 3 | 1 | 42 | 5 | No |
| 4 | 5 | 134 | 56 | Yes |
| 5 | 2 | 62 | 16 | Yes |
| 6 | 2 | 84 | 24 | Yes |
| 7 | 2 | 31 | 24 | No |
| 8 | 4 | 120 | 45 | Yes |

The first participant did not have a mood improvement had a perfect score of 12, which is higher than the average of the participants that had mood improvement, but only had a time of 15 seconds for a single round which is 51.9% of the average time it took for participants with an improvement in mood. After asking for feedback, the first participant was rushed and tried to finish the game as fast as possible, leading to a slight increase in stress. The third participant took 42 seconds to play one round, 145% of the average time it took for participants with an improvement in mood, and got a score of 5. The third participant wrote that they did not understand the rules very well and spent most of their time figuring out how to play. The seventh participant was very similar to the first participant. The seventh participant played two rounds and spent an average of 15.5 seconds per round and a perfect score on both rounds. As feedback, they said the game was too easy and it wasn't fun to play.

Of the participants that did have an improvement in mood, some of the common responses to the question of why they had an improvement in mood included the music and the calming repetition of clicking and dragging in addition to the fact that the matching pairs were very obvious to spot.

## 5.2. Experiment 2: How can we determine if the game is successful in stress relief without having to ask the user?

We used different machine learning models to determine the accuracy with which we can predict whether a user's stress will be relieved. We used a Linear Kernel Support Vector Machine, a Random Forest Classifier, a Logistic Regression Model, and a Gradient Boosting classifier. These models were trained using 500 pieces of dummy data that were randomly generated. Each model was tested five times and the average of the five scores was used to determine which model is most accurate in its predictions. The Random Forest Classifier had a max depth of 10 and a random state of 0, the Logistic Regression model had a random state of 0, and the Gradient Boosting Classifier had 100 n estimators, a learning rate of 1.0, a max depth of 5, and a random state of 0.

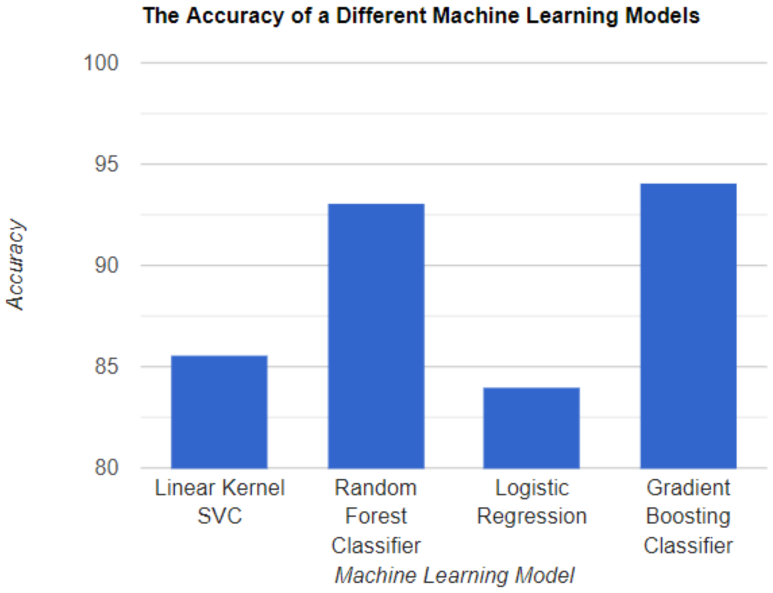


Figure 7. The accuracy of a different machine learning models

Of the four models we tested, the Gradient Boosting Classifier had the highest average accuracy at 94.8% accuracy followed by the Random Forest Classifier with an accuracy of 93.6%. The Linear Kernal SVC had an accuracy of 85.6% while the Logistic Regression model had an accuracy of 84%. This shows that our problem is not a logistic relationship. We wanted to get the highest accuracy possible, so we decided to test a few parameters in the Gradient Boosting Classifier.

### 5.3. Experiment 3: How do the parameters of the Gradient Boosting Classifier affect our accuracy?

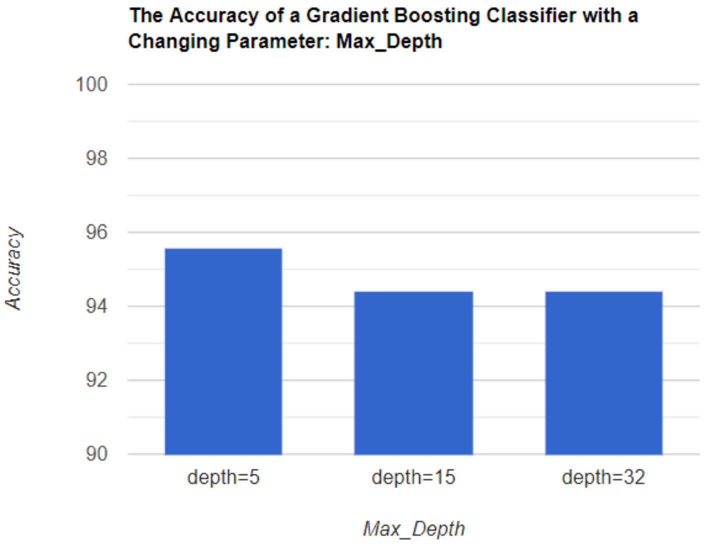


Figure 8. The accuracy of a gradient boosting classifier (1)

The Gradient Boosting Classifier had a parameter called max_depth that adjusted the amount of leaves on the learning tree. We tested the Gradient Boosting Classifier using the depths of 5, 15, 32. The Gradient Boosting Classifier had an accuracy of 95.6% accuracy at a max depth of 5 and 94.4% at max depth of 15 and 32. This is contradictory to what is expected, because having more leaves on the learning tree usually leads to having a higher accuracy. This unexpected result could be because our pattern is relatively simple and having a higher number of leaves on the learning tree was unnecessary.

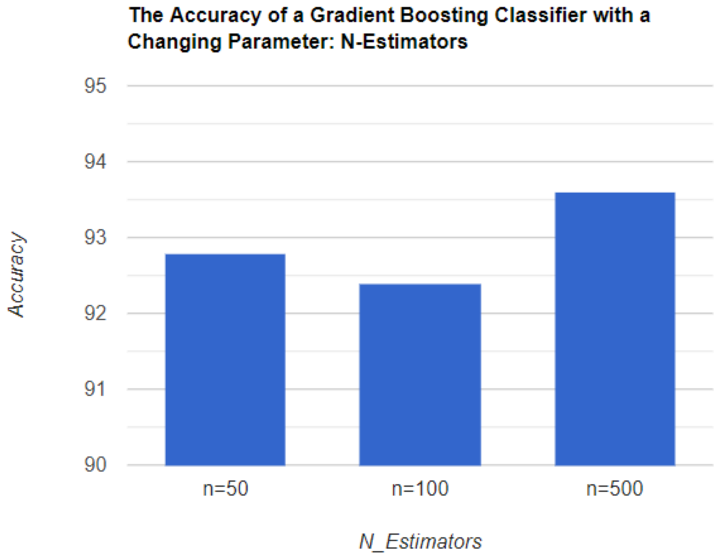


Figure 9. The accuracy of a gradient boosting classifier (2)

When adjusting the n estimators, the Gradient Boosting Classifier had an accuracy of 92.8% at n=50, 92.4% at n=100, and 93.6% at n=500. While adjusting both of these parameters had an impact on accuracy, the differences in accuracy were very small. We did, however, change our algorithm to have 500 n estimators instead of the original 100.

## 5.4. Experiment 4: How does the size of a data set affect the accuracy of the Gradient Boosting Classifier?

Changing the data set size also affected the accuracy of our experiment. When only using 50 pieces of data, we had an accuracy of 84%. We had an accuracy of 82% with 100 pieces of data, 92.8% with 250 pieces of data, and 96.4% with 500 pieces of data. The average of 100 pieces of data was lower than expected. With the exception of the experiment with 100 pieces of data, the general trend is that the larger size of the data set, the more accurate the Gradient Boosting Classifier will be.

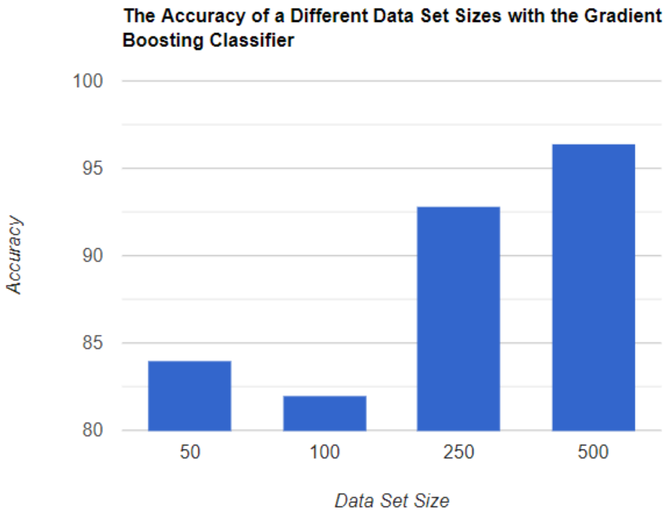


Figure 10. The accuracy of a different data set sizes

### 5.5. Experiment 5: How does the cross evaluation test ratio affect our accuracy?

We had an accuracy of 95.2% with a cross evaluation test size of 0.1, which is the test size we used for our previous experiments. Using a cross evaluation test size of 0.5, we got an accuracy of 90.16%. We got an accuracy of 100% with the test size of 1. Of the tested values, having a test size of 1 is not ideal, because a machine learning algorithm having a 100% accuracy is highly unlikely.

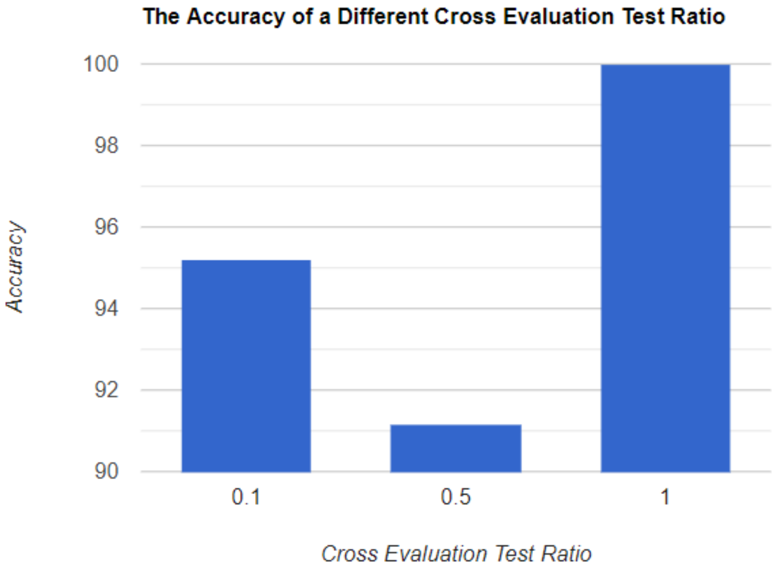


Figure 11. The accuracy of a different cross evaluation test ratio

## 5.6. Analysis of Results

Our game seems to relieve the user's stress if the user takes the time to play the game through without rushing and if the user understands the rules of the game. The second condition is a gaming standard since many users would not be able to enjoy any game if they did not understand the rules even if they were not penalized for it. In the end, we were able to develop an algorithm that was able to relatively accurately predict whether a user would have benefitted from the game without having to ask them directly. We decided to use the Gradient Booster Classifier as opposed to any other algorithm because it gave us the highest accuracy at default settings. In order to increase the accuracy higher, we adjusted the max depth, the number of n estimators, data set size, and the cross evaluation test ratio. We found that we had the highest reasonable accuracy with a max depth of 5, 500 n estimators, 500 pieces of data, and a cross evaluation ratio of 0.1.

# 6. CONCLUSIONS

We created a game for the purpose of relaxation and de-stressing using different minigames including: a journal minigame where users are encouraged to reflect upon their emotions, a matching minigame where users will be able to connect with NPCs, and a flower popping minigame where users will repeat a clicking motion for relaxation. All three of these minigames provide an incentive for success so the user will continue to play the game. In order to find out which minigame is most effective in destressing its users, we asked many different people to rate the games [13].

Our application may not be as aesthetically pleasing as some other games, which may have a large factor in determining the ability for users to fully enjoy the game. The idea itself seems to be very entertaining, which means that as the game is developed further, it is possible for it to become even more successful in relieving stress. However, the most successful machine learning algorithm, the Linear Kernal SVC, also has an accuracy only a bit better than random chance. There is also still nowhere that this prediction is put into play.

In the future, we plan on placing more weight on certain factors in the machine learning algorithm so we will be able to determine whether a person's mood improved without having to ask them for the second time before they close the application. Once this has been improved, we may be able to implement a function that encourages users to continue playing if their mood has not been improved.

## REFERENCES

[1] Jaycox, Lisa H., et al. "Impact of teen depression on academic, social, and physical functioning." Pediatrics 124.4 (2009): e596-e605.

[2] Cast, Alicia D., and Peter J. Burke. "A theory of self-esteem." Social forces 80.3 (2002): 1041-1068.

[3] Prince, Martin, et al. "No health without mental health." The lancet 370.9590 (2007): 859-877.

[4] Hodkinson, Paul. "Interactive online journals and individualization." New Media & Society 9.4 (2007): 625- 650.

[5] Mondesire, Sean, and R. Paul Wiegand. "Evolving a non-playable character team with layered learning." 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM). IEEE, 2011.

[6] Brunyé, Tad T., et al. "Learning to relax: Evaluating four brief interventions for overcoming the negative emotions accompanying math anxiety." Learning and Individual Differences 27 (2013): 1-7.

[7] Perou R, Bitsko RH, Blumberg SJ, Pastor P, Ghandour RM, Gfroerer JC, Hedden SL, Crosby AE, Visser SN, Schieve LA, Parks SE, Hall JE, Brody D, Simile CM, Thompson WW, Baio J, Avenevoli S, Kogan MD, Huang LN. Mental health surveillance among children – United States, 2005—2011.

MMWR 2013;62(Suppl; May 16, 2013)

[8] Son, Changwon et al. “Effects of COVID-19 on College Students' Mental Health in the United States: Interview Survey Study.” Journal of medical Internet research vol. 22,9 e21279. 3 Sep. 2020, doi:10.2196/21279

[9] Ma ł gorzata, Plechawska-W ó jcik, and Jakub Grzesiak. "IMPLEMENTATION OF SERIOUS GAME TECHNIQUES IN RAISING THE SOCIAL AWARENESS OF THE DEPRESSION DISEASE."

[10] Poels, Karolien, Yvonne De Kort, and Wijnand IJsselsteijn. "Identification and categorization of digital game experiences: a qualitative study integrating theoretical insights and player perspectives." Westminster Papers in Communication and Culture 9.1 (2012): 107-129.

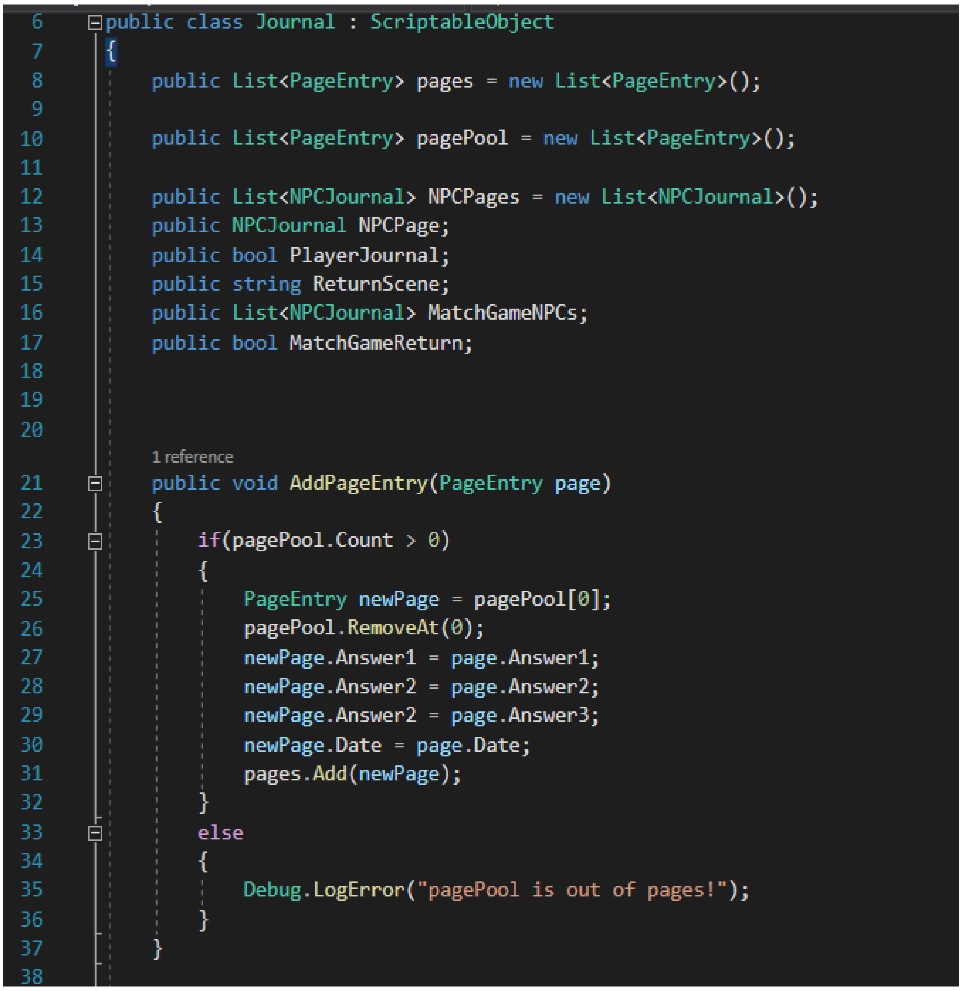
[11] Mercer, Nicole. "Stress Relieving Video Games: Creating a Game for the Purpose of Stress Relief and Analyzing Its Effectiveness." (2015).

[12] Nacke, Lennart Erik, et al. "Biofeedback game design: using direct and indirect physiological control to enhance game interaction." Proceedings of the SIGCHI conference on human factors in computing systems. 2011.

[13] Konicek, Petr. "Destressing." Rockburst. Butterworth-Heinemann, 2018. 453-471.

# Appendix A

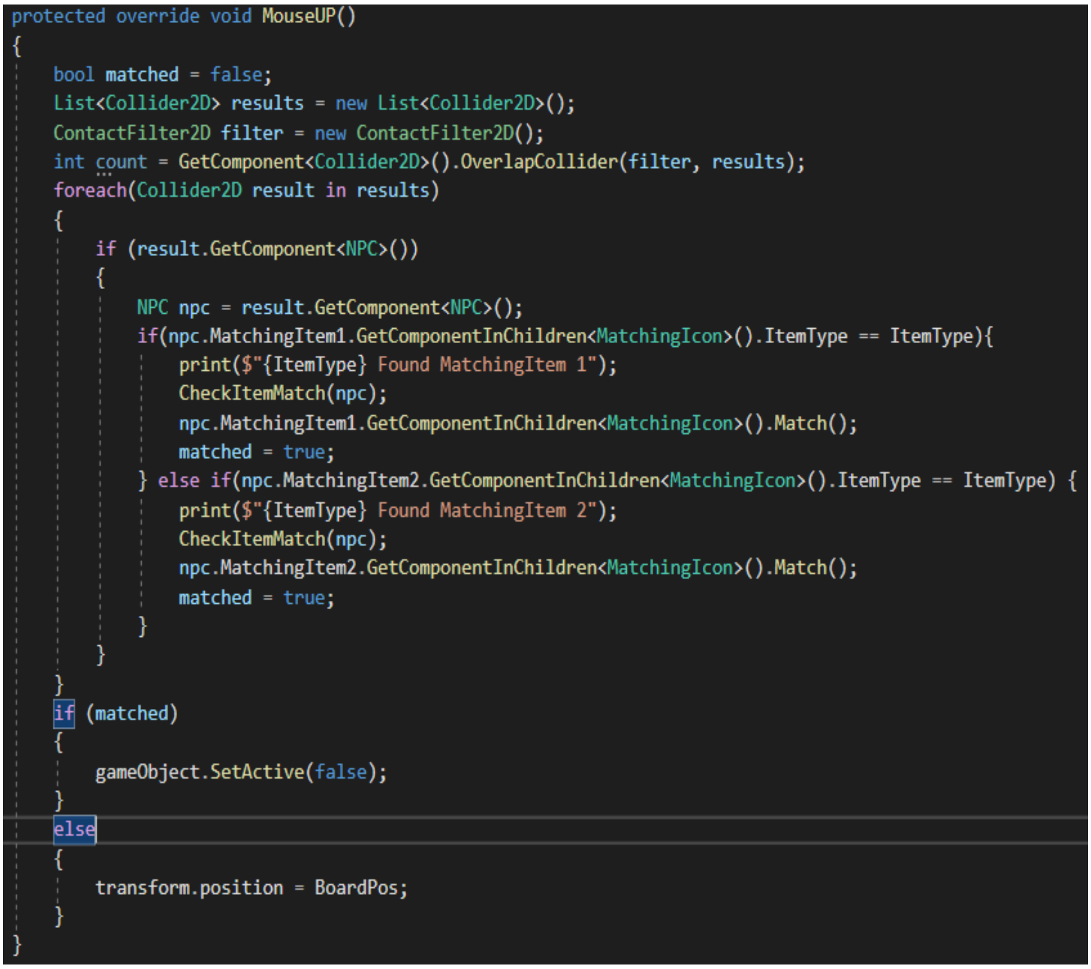
Code segment of Minigame 1



```
public class Journal : ScriptableObject
{
    public List<PageEntry> pages = new List<PageEntry>();

    public List<PageEntry> pagePool = new List<PageEntry>();

    public List<NPCJournal> NPCPages = new List<NPCJournal>();
    public NPCJournal NPCPage;
    public bool PlayerJournal;
    public string ReturnScene;
    public List<NPCJournal> MatchGameNPCs;
    public bool MatchGameReturn;



    public void AddPageEntry(PageEntry page)
    {
        if(pagePool.Count > 0)
        {
            PageEntry newPage = pagePool[0];
            pagePool.RemoveAt(0);
            newPage.Answer1 = page.Answer1;
            newPage.Answer2 = page.Answer2;
            newPage.Answer2 = page.Answer3;
            newPage.Date = page.Date;
            pages.Add(newPage);
        }
        else
        {
            Debug.LogError("pagePool is out of pages!");
        }
    }

```

# Appendix B

Code segment of Minigame 2



```
protected override void MouseUP()
{
    bool matched = false;
    List<Collider2D> results = new List<Collider2D>();
    ContactFilter2D filter = new ContactFilter2D();
    int count = GetComponent<Collider2D>().OverlapCollider(filter, results);
    foreach(Collider2D result in results)
    {
        if (result.GetComponent<NPC>())
        {
            NPC npc = result.GetComponent<NPC>();
            if(npc.MatchingItem1.GetComponentInChildren<MatchingIcon>().ItemType == ItemType){
                print($"{ItemType} Found MatchingItem 1");
                CheckItemMatch(npc);
                npc.MatchingItem1.GetComponentInChildren<MatchingIcon>().Match();
                matched = true;
            } else if(npc.MatchingItem2.GetComponentInChildren<MatchingIcon>().ItemType == ItemType) {
                print($"{ItemType} Found MatchingItem 2");
                CheckItemMatch(npc);
                npc.MatchingItem2.GetComponentInChildren<MatchingIcon>().Match();
                matched = true;
            }
        }
    }
    if (matched)
    {
        gameObject.SetActive(false);
    }
    else
    {
        transform.position = BoardPos;
    }
}
```

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN AUTOMATED ANALYTICS ENGINE FOR COLLEGE PROGRAM SELECTION USING MACHINE LEARNING AND BIG DATA ANALYSIS

Jinhui Yu, Xinyu Luan and Yu Sun

California State Polytechnic University, Pomona, CA, 91768

## *ABSTRACT*

*Because of the differences in the structure and content of each website, it is often difficult for international applicants to obtain the application information of each school in time. They need to spend a lot of time manually collecting and sorting information. Especially when the information of the school may be constantly updated, the information may become very inaccurate for international applicants. we designed a tool including three main steps to solve the problem: crawling links, processing web pages, and building my pages. In compiling languages, we mainly use Python and store the crawled data in JSON format [4]. In the process of crawling links, we mainly used beautiful soup to parse HTML and designed crawler.*

*In this paper, we use Python language to design a system. First, we use the crawler method to fetch all the links related to the admission information on the school's official website. Then we traverse these links, and use the noise_remove [5] method to process their corresponding page contents, so as to further narrow the scope of effective information and save these processed contents in the JSON files. Finally, we use the Flask framework to integrate these contents into my front-end page conveniently and efficiently, so that it has the complete function of integrating and displaying information.*

## *KEYWORDS*

*Data Crawler, Data Processing, Web framework.*

## 1. INTRODUCTION

International students will encounter many problems when they apply for American universities and collecting enrollment information is absolutely one of them. When we tried to find the admission information of several schools, we found that it would take me a lot of time to find information one by one. As far as we know, every year, millions of students apply to American universities from all over the world, which is with no doubt a significant part of American education system. To get accurate and complete information about their dream schools, students have to browse websites of lots of universities to seek for valuable information. However, the structures of these websites vary greatly, and majors can be categorized in totally different ways, which make it easy for students to get lost in the huge amount of information. Students' time is precious, and they should be able to get information they need easily and efficiently.

Although there are already some websites and apps to help international applicants integrate their information, these sites are unable to update information in time due to certain technical

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 193-205, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111417

problems, which may adversely affect the application process. The data of these web pages is often manually entered when

programmers make web pages. Obviously, such a huge amount of data cannot be changed in time manually. This paper designs a tool to crawl the admissions data from various schools and visualize these information in a systematic and concise way which could help students shorten decision making time.

In terms of development language, PHP, Java and C are the mainstream programming languages, but they all have some problems. PHP is known as "the best language in the world" [6], but it has no concept of multithreading, not much support for asynchrony and so on, which are its shortcomings as a crawler: Although C language and C + + are the most efficient and performance languages, but the learning cost is very high, and the amount of code is too large. Java has a complete ecosystem and is the biggest competitor of python, but its code is very cumbersome, and the crawler needs to modify the code frequently.

Many methods can be used to parse the captured web pages; both beautifulsoup [7] and lxml [8] are important tools in web page parsing and information extraction. Lxml is a library written in Python, which can deal with XML and HTML quickly and flexibly. Although it has higher performance in dealing with HTML, its learning cost is relatively high, so it is not suitable for small projects.

In terms of web framework, Django is the most versatile Web development framework in the Python field, with complete functions, good maintainability and development speed, but it will be cumbersome for small project code. Tornado is a web server and web application framework written in Python language. Although it has fast processing speed, it has few plug-ins and is not very convenient and efficient.

My goal is to make a crawler program to get the admission information of major universities. Web crawler (also known as web spider) is a program that automatically grabs the information of the World Wide Web according to certain rules. There are some good methods we have used to build the system.

Firstly, we chose Python for development, because Python has high portability, and has a very powerful third-party library. The development based on this can greatly improve the development efficiency. And compared with Java and C language, Python code is more concise and easy to read and write.

Second, we use BeautifulSoup when we process a web page after crawling. BeautifulSoup is a Python library that fetches data from web pages. The advantage of BS is that, compared with Scrapy crawler framework, although the amount of code will be increased, it allows me to crawl information more flexibly and freely.

Third, in order to build the web to show the information we integrate, we chose the Flask framework, a lightweight and customizable web framework [1]. Flask has some obvious advantages. First of all, compared with Django and other similar frameworks, flask is more flexible, light, safe and easy to use. It allows me to complete the implementation of small and medium-sized websites or web services with rich functions in a short time. In addition, flask also has strong customization. we can add functions according to my own needs, and realize the enrichment and expansion of functions while keeping the core functions simple. Its powerful plug-in library allows me to realize personalized website customization and develop a powerful website.

The rest of the paper is organized as follows: Section 2 details the challenges we encountered during the experiment and design of the sample; Section 3 focuses on the details of the solution corresponding to the challenges we mentioned in Section 2; followed by section 4 which describes the related work and finally section 5 which provides a concluding comment and points out the future work of the project.

## 2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

### 2.1. Finding Right Link to Use

The first problem we encountered was that when we started crawling with the admission page of a school's official website as the source page, we was frustrated that the results contained too many irrelevant URLs and even many external links. This is because each school's official website contains news, notices, events and other information. we need to determine which pages are related to application and enrollment, and crawl the links in these pages; and these pages sometimes contain some extra information. Links, the pages corresponding to these links are basically irrelevant to the application information, so we don't need to save these links. So we need to design a method to remove useless links and save all the links related to the application information so that we can crawl the content of the corresponding pages after these links.

### 2.2. Finding Out Useful Information

The second challenge that bothers me is that the crawled pages often include a lot of content, such as navigation bars, links, pictures, videos, texts, etc., all in the form of tags in the code. These irrelevant content are like "noise", making it difficult for me to quickly find the main content and information. How can we find the information we need from so many elements? When we crawled the data from the school's official website, we found that the code contains a lot of CSS and other elements. This part of the code is mainly used to modify the web page, such as adjusting the layout of the web page, font, color, pictures, etc., to make the display of the web page look better and regular, but they are not important. So we designed the method, we first remove some common modifier tags, so that we can locate the key tags and the text content in them more quickly. we then determine the position of the content block by calculating the number of tags, so as to obtain the main content of the page.

## 3. SOLUTION

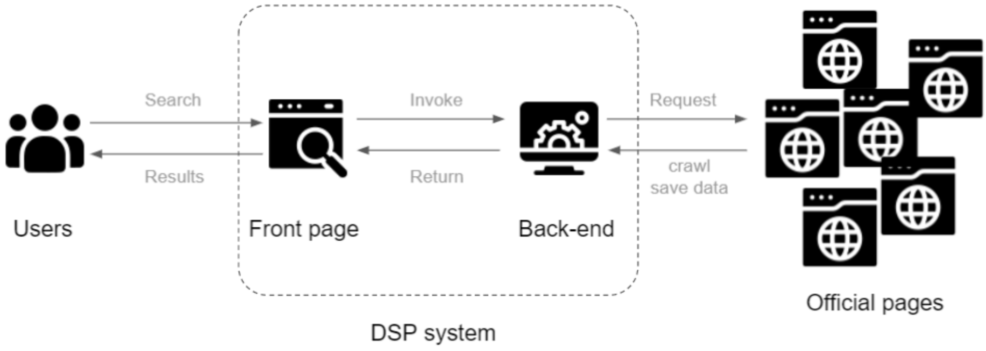


Figure 1. The overview of DSP system

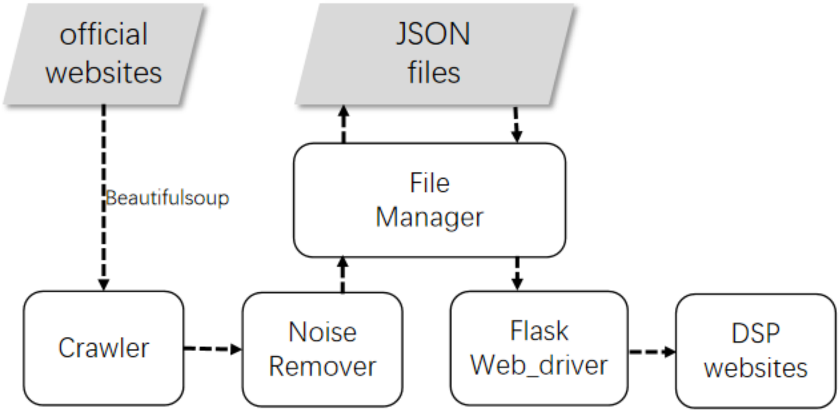


Figure 2. The o

DSP is a system based on crawler technology to assist students to apply the most well-known universities in the United States. DSP integrates the information of those programs. Users can browse the important information including university ranking, application conditions, department courses and so on; At the same time, users can also search for the Enrollment Requirements of various colleges and universities by searching professional names and courses, so as to quickly prepare for applying for universities. In DSP, we first crawl and save the enrollment information on the official websites of colleges and universities or some American authoritative websites, then filter, classify and integrate these information and data, and finally present them on our website in a more concise and easy way. Therefore, DSP can present the enrollment information of hundreds of well-known universities in the United States. Because all the information in DSP is obtained by crawling the official websites of universities, it can basically keep pace with the official websites of universities and update in real time, which means it will not provide applicants with expired or wrong information. Because the data processed by DSP will only contain enrollment information, applicants can use it to easily and quickly find the admission requirements, application deadline, tuition fees and other application related information of their favorite colleges, without being disturbed by other irrelevant content. The main technical difficulty of implementing a DSP system is how to automatically crawl and classify the enrollment information, since the structure of different official websites is very different. In addition, the processed data must be put on the front page in a more unified way, which is convenient for applicants to query. To achieve these goals, our tool consists of 5 main components (see the thick boxes in Figure 2):

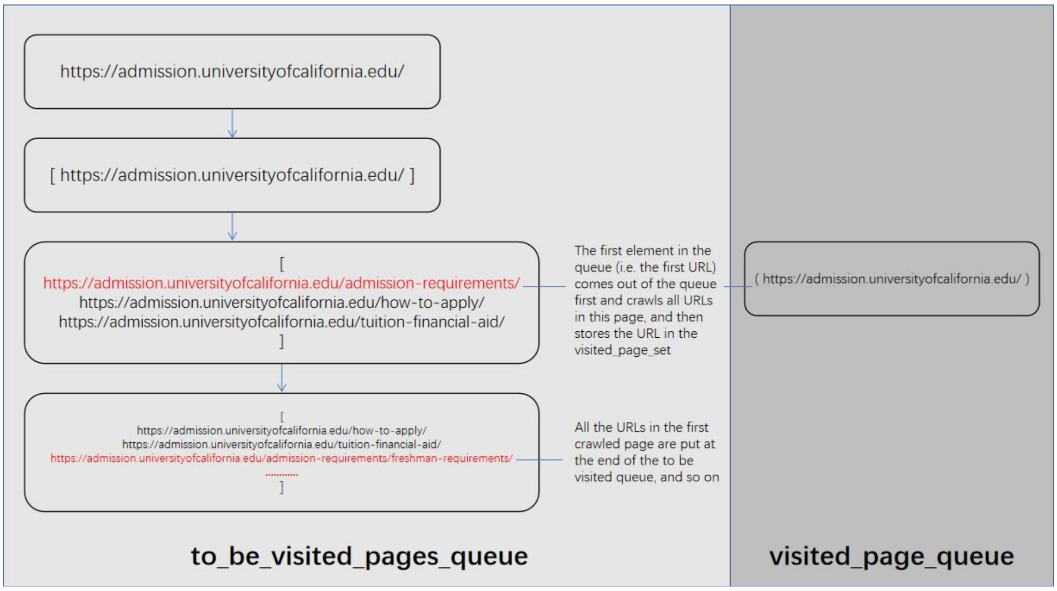
## Crawler

The most important goal of the whole DSP function is that we need to obtain all the links in the application or admission pages of the official websites of various schools, and store them in a container, so that we can crawl the contents of the corresponding pages of these links in the next step. In the process of crawling pages, DSP needs to solve two problems: duplicated pages removal and ensuring crawling within a certain range.

In the early process of crawling pages [9], we found that the links of a web page in a website had a loop. For example, we could see the link of the home page in the home page of the website, and then we might see a link to the home page in the sub-pages, and perhaps the sub-pages would also have corresponding links to the home page. Such crawling will lead to repeated crawling of the web page. To solve this problem, we need to design a class called Noise_remover to remove the duplicated pages.

In this class, my major idea was using Breadth First Search (BFS)[10] to crawl the URL and then processing it. In BFS, we need to use a data structure called queue. Queue [11] is a linear storage table, in which the element data is inserted at one end of the table and deleted at the other end, thus forming a FIFO (First In First Out) [12] table. So we defined a queue to_be_visited_pages_queue) to store the URLs to be crawled, and a set(visited_page_set) to store the crawled URLs. Set is a mutable container in Python whose data is not repeatable.

In Breadth First Search, we took the first element in the queue to crawl, then put all the URLs in the next layer at the end of the queue, and saved the visited URLs in the set. So in the next round of crawling, we could judge whether the URL was already visited_ page_ set to remove pages which we had already crawled.

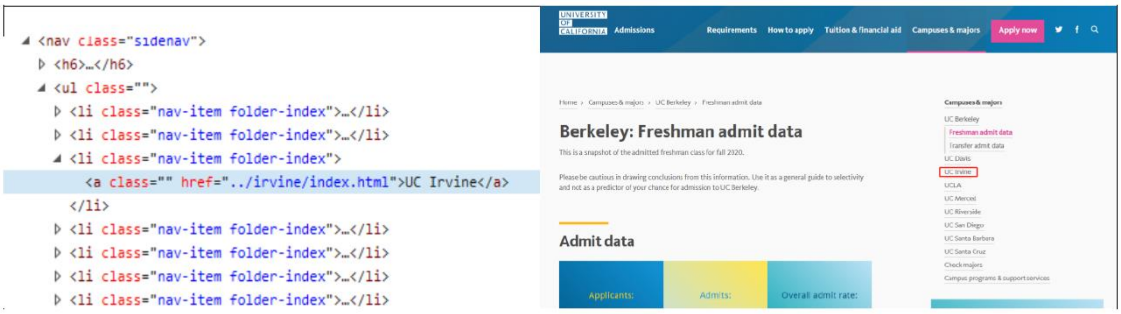

https://admission.universityofcalifornia.edu/
[ https://admission.universityofcalifornia.edu/ ]
[
https://admission.universityofcalifornia.edu/admission-requirements/
https://admission.universityofcalifornia.edu/how-to-apply/
https://admission.universityofcalifornia.edu/tuition-financial-aid/
]
[
https://admission.universityofcalifornia.edu/how-to-apply/
https://admission.universityofcalifornia.edu/tuition-financial-aid/
https://admission.universityofcalifornia.edu/admission-requirements/freshman-requirements/
............
]
( https://admission.universityofcalifornia.edu/ )
to_be_visited_pages_queue
visited_page_queue

The first element in the queue (i.e. the first URL) comes out of the queue first and crawls all URLs in this page, and then stores the URL in the visited_page_set

All the URLs in the first crawled page are put at the end of the to be visited queue, and so on

Figure 3. The o

In addition, when we get the links in the page, we find that many of these links do not appear as complete links, but in the form of "../irvine/index/html". So we must complete their links before crawling.



```
◢ <nav class="sidenav">
  ▷ <h6>…</h6>
  ◢ <ul class="">
    ▷ <li class="nav-item folder-index">…</li>
    ▷ <li class="nav-item folder-index">…</li>
    ◢ <li class="nav-item folder-index">
        <a class="" href="../irvine/index.html">UC Irvine</a>
      </li>
    ▷ <li class="nav-item folder-index">…</li>
    ▷ <li class="nav-item folder-index">…</li>
    ▷ <li class="nav-item folder-index">…</li>
    ▷ <li class="nav-item folder-index">…</li>
```

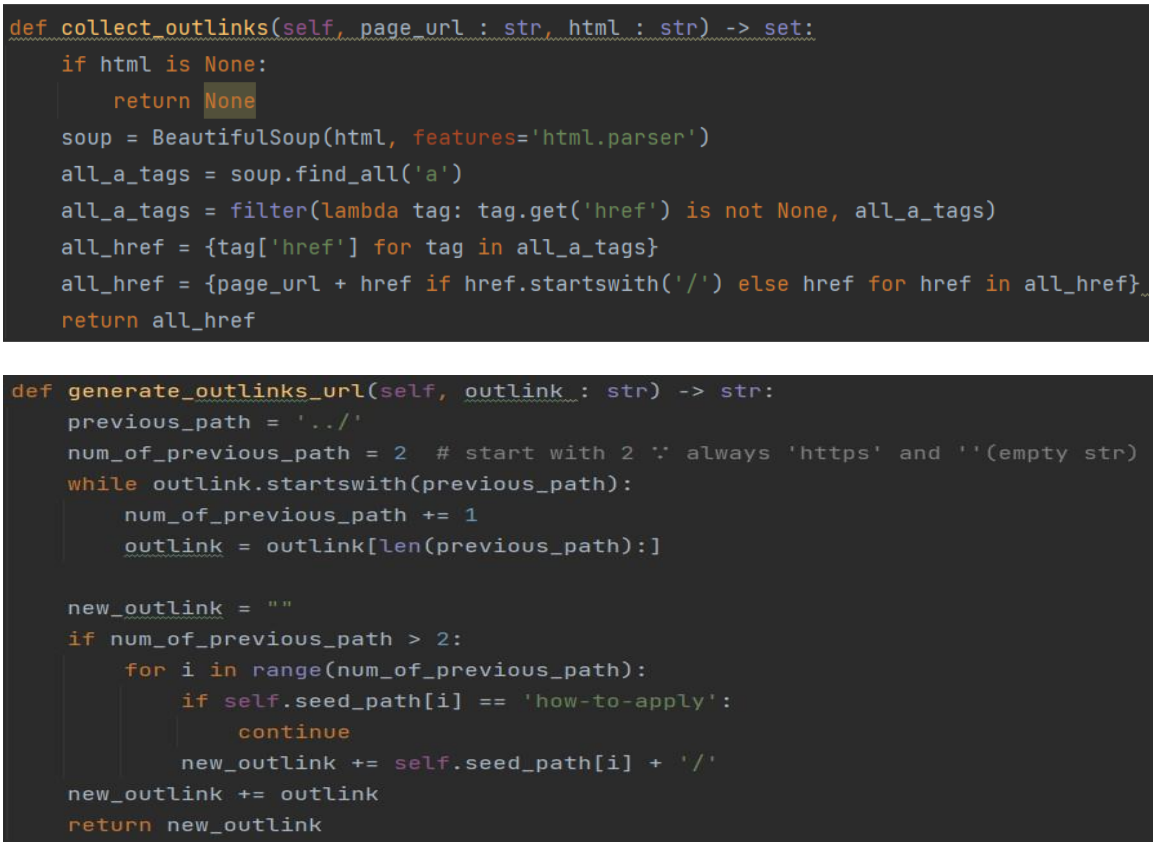
Berkeley: Freshman admit data

This is a snapshot of the admitted freshman class for fall 2020.

Please be cautious in drawing conclusions from this information. Use it as a general guide to selectivity and not as a predictor of your chance for admission to UC Berkeley.

Admit data

I first used the function called collect_outlinks to get all URLs in a page, no matter whether they were complete or not. Then, we used the function called generate_outlinks_url to judge whether the URL was complete or not. If the URL was not complete, it would be automatically completed.

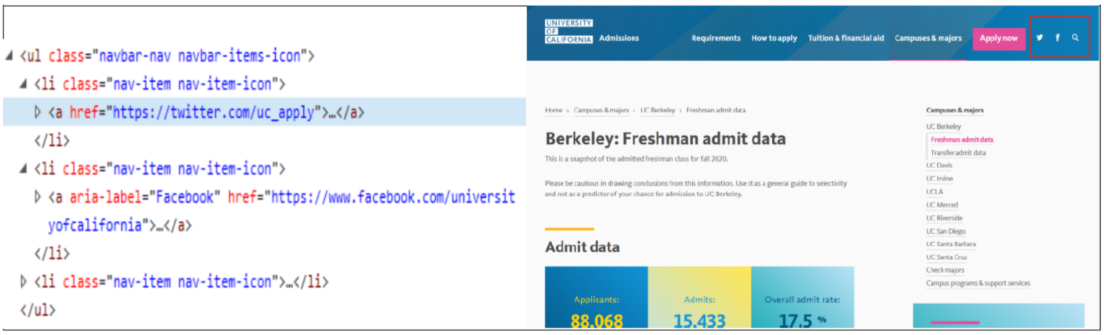


```python
def collect_outlinks(self, page_url : str, html : str) -> set:
    if html is None:
        return None
    soup = BeautifulSoup(html, features='html.parser')
    all_a_tags = soup.find_all('a')
    all_a_tags = filter(lambda tag: tag.get('href') is not None, all_a_tags)
    all_href = {tag['href'] for tag in all_a_tags}
    all_href = {page_url + href if href.startswith('/') else href for href in all_href}
    return all_href
```

```python
def generate_outlinks_url(self, outlink : str) -> str:
    previous_path = '../'
    num_of_previous_path = 2  # start with 2 ∵ always 'https' and ''(empty str)
    while outlink.startswith(previous_path):
        num_of_previous_path += 1
        outlink = outlink[len(previous_path):]

    new_outlink = ""
    if num_of_previous_path > 2:
        for i in range(num_of_previous_path):
            if self.seed_path[i] == 'how-to-apply':
                continue
            new_outlink += self.seed_path[i] + '/'
    new_outlink += outlink
    return new_outlink
```

Figure 4. The o

There was still a small issue to be addressed. When crawling the school websites, we hope all the URLs are the internal pages of the schools, not some external links. As shown in the figure below, the page contains some links to Twitter and Facebook, which we need to remove as well.



```html
◢ <ul class="navbar-nav navbar-items-icon">
  ◢ <li class="nav-item nav-item-icon">
    ▷ <a href="https://twitter.com/uc_apply">…</a>
    </li>
  ◢ <li class="nav-item nav-item-icon">
    ▷ <a aria-label="Facebook" href="https://www.facebook.com/universit
      yofcalifornia">…</a>
    </li>
  ▷ <li class="nav-item nav-item-icon">…</li>
  </ul>
```

Berkeley: Freshman admit data

This is a snapshot of the admitted freshman class for fall 2020.

Please be cautious in drawing conclusions from this information. Use it as a general guide to selectivity and not as a predictor of your chance for admission to UC Berkeley.

Admit data

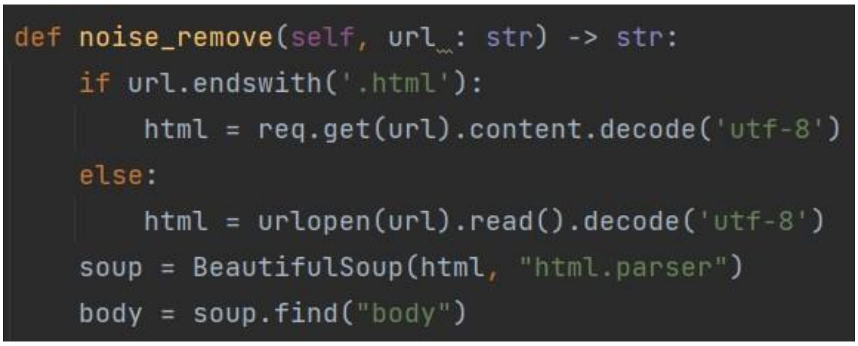
Applicants: 88,068 | Admits: 15,433 | Overall admit rate: 17.5 %

Figure 5. The o

## Noise_ remover

After fetching all the links to admission related pages, we need to get more information about the page and the application, so we designed a class called Noise_ remover.

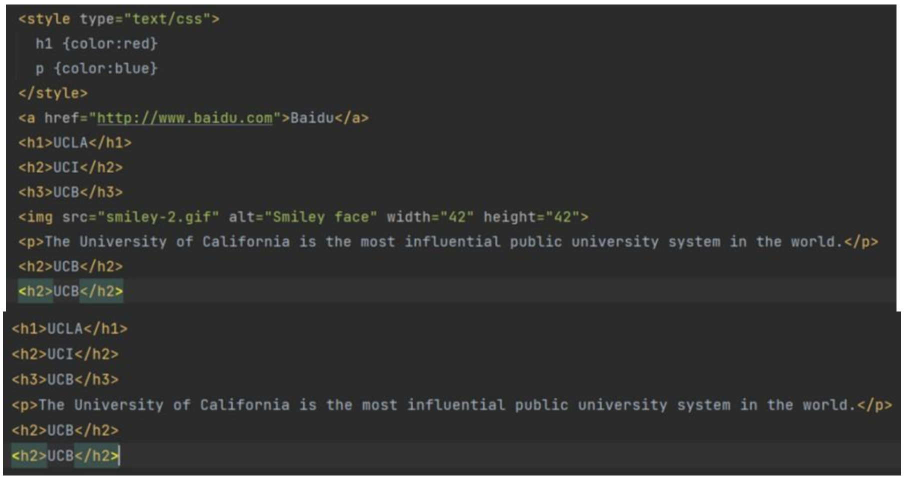
In Noise_ remover, we mainly use the noise_ remove function. This function first obtained the web page content through requests generating HTML data, and then used Beautifulsoup to parse the HTML data. Beatifulsoup is a kind of HTML parser, which uses knowledge of the syntax of the markup language to identify the structure [1]. At the same time, it provides easy-to-use navigation, search and modification operations, which greatly saves my programming time.



```
def noise_remove(self, url : str) -> str:
    if url.endswith('.html'):
        html = req.get(url).content.decode('utf-8')
    else:
        html = urlopen(url).read().decode('utf-8')
    soup = BeautifulSoup(html, "html.parser")
    body = soup.find("body")
```

Figure 6. The o

After generating the beautiful soup object, we first need to remove the tags used to decorate the web page, such as CSS. we find the body tag through the built-in find method of beautiful soup, and then use the select function to find some of the CSS tags, such as body. Select ("image"), and then use the extract function to remove the modificatory content, so that only the main content of a page is left in the body tag.

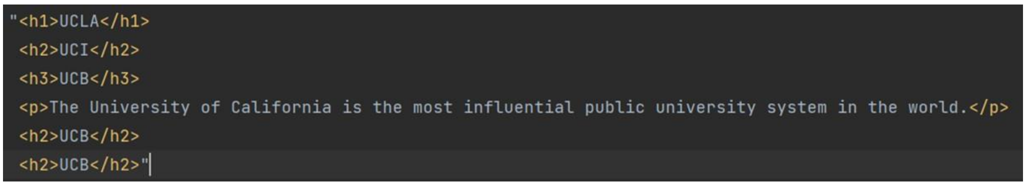


```
<style type="text/css">
  h1 {color:red}
  p {color:blue}
</style>
<a href="http://www.baidu.com">Baidu</a>
<h1>UCLA</h1>
<h2>UCI</h2>
<h3>UCB</h3>
<img src="smiley-2.gif" alt="Smiley face" width="42" height="42">
<p>The University of California is the most influential public university system in the world.</p>
<h2>UCB</h2>
<h2>UCB</h2>
```

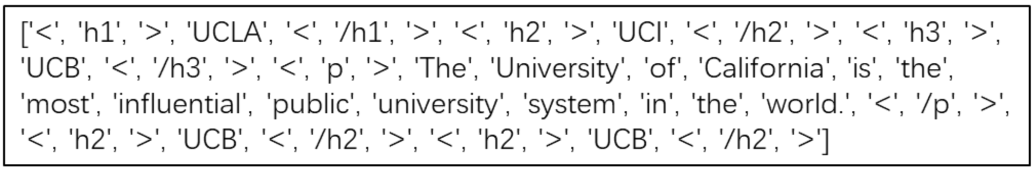
```
<h1>UCLA</h1>
<h2>UCI</h2>
<h3>UCB</h3>
<p>The University of California is the most influential public university system in the world.</p>
<h2>UCB</h2>
<h2>UCB</h2>
```

Figure 7. Before After

I imported NLTK package [13] (Natural Language Toolkit), which is a python library commonly used in the NLP research field. First, we use str () to convert the content of <body> from BS4. Element. Tag type to string type; Next, we use word_ tokenize function to cut the string into words one by one.



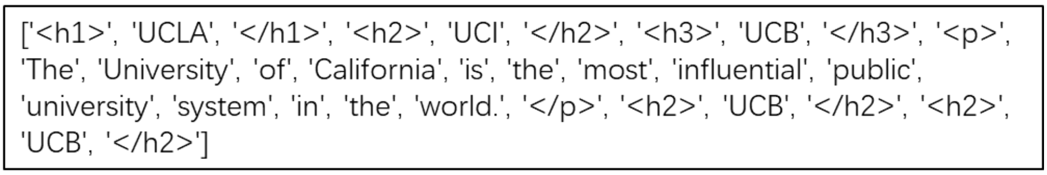
```
"<h1>UCLA</h1>
 <h2>UCI</h2>
 <h3>UCB</h3>
 <p>The University of California is the most influential public university system in the world.</p>
 <h2>UCB</h2>
 <h2>UCB</h2>"
```



```
['<', 'h1', '>', 'UCLA', '<', '/h1', '>', '<', 'h2', '>', 'UCI', '<', '/h2', '>', '<', 'h3', '>',
'UCB', '<', '/h3', '>', '<', 'p', '>', 'The', 'University', 'of', 'California', 'is', 'the',
'most', 'influential', 'public', 'university', 'system', 'in', 'the', 'world.', '<', '/p', '>',
'<', 'h2', '>', 'UCB', '<', '/h2', '>', '<', 'h2', '>', 'UCB', '<', '/h2', '>']
```

Figure 8. The o

Then we designed a customize_ tokenizer function, traverses the cut content, and reassembles "<" and ">" and the content between them into a content list.



```
['<h1>', 'UCLA', '</h1>', '<h2>', 'UCI', '</h2>', '<h3>', 'UCB', '</h3>', '<p>',
'The', 'University', 'of', 'California', 'is', 'the', 'most', 'influential', 'public',
'university', 'system', 'in', 'the', 'world.', '</p>', '<h2>', 'UCB', '</h2>', '<h2>',
'UCB', '</h2>']
```

Figure 9. The o

Now, there are still a lot of useless tags in the list. To find content blocks, we count the amount of tags to analyze cumulative distribution of tags in the targeted pages.

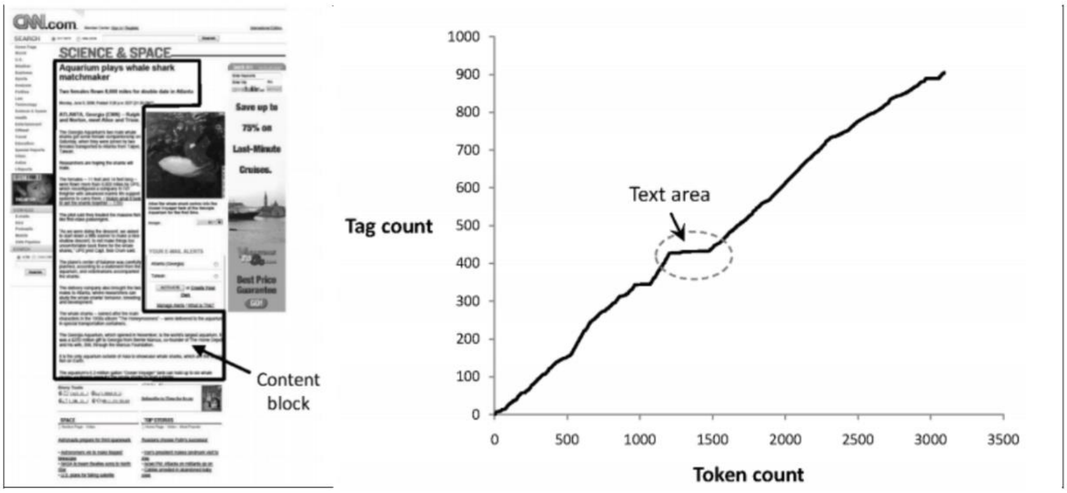
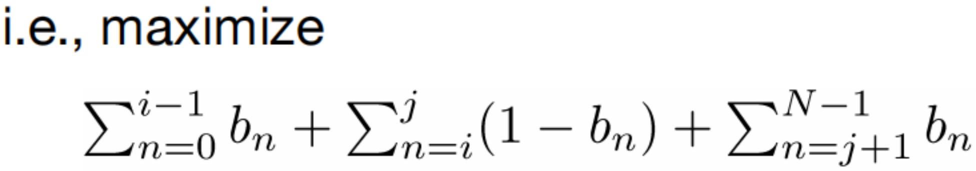


Figure 10. Noise Remove [2]

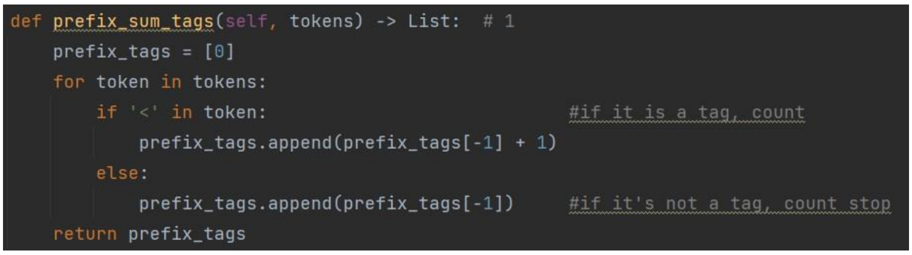
Now, there are still a lot of useless tags in the list. To find content blocks, we need to count the amount of tags to analyze the cumulative distribution of tags in the targeted pages, and the main text content of the page corresponds to the “plateau” in the middle of the distribution. This flat area is relatively small because of the large amount of formatting and presentation information in the HTML source for the page[2]. So we design a method to determine where the “plateau” is. Represent a web page as a sequence of bits, where bn = 1 indicates that the nth element in the content list is a tag. The values of we and j which maximize both the number of tags below we and above j and the number of non-tag tokens between we and j can help me find the position of main text content. we designed a function called prefix_ sum_ Tags , in which we use a new list to count.



i.e., maximize

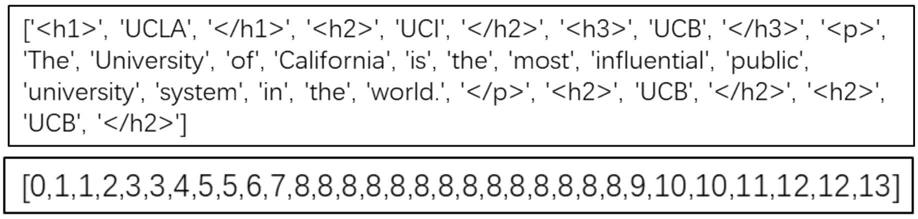
$$\sum_{n=0}^{i-1} b_n + \sum_{n=i}^{j}(1-b_n) + \sum_{n=j+1}^{N-1} b_n$$

Figure 11. The o



```
def prefix_sum_tags(self, tokens) -> List:  # 1
    prefix_tags = [0]
    for token in tokens:
        if '<' in token:                           #if it is a tag, count
            prefix_tags.append(prefix_tags[-1] + 1)
        else:
            prefix_tags.append(prefix_tags[-1])    #if it's not a tag, count stop
    return prefix_tags
```

Figure 12: The o



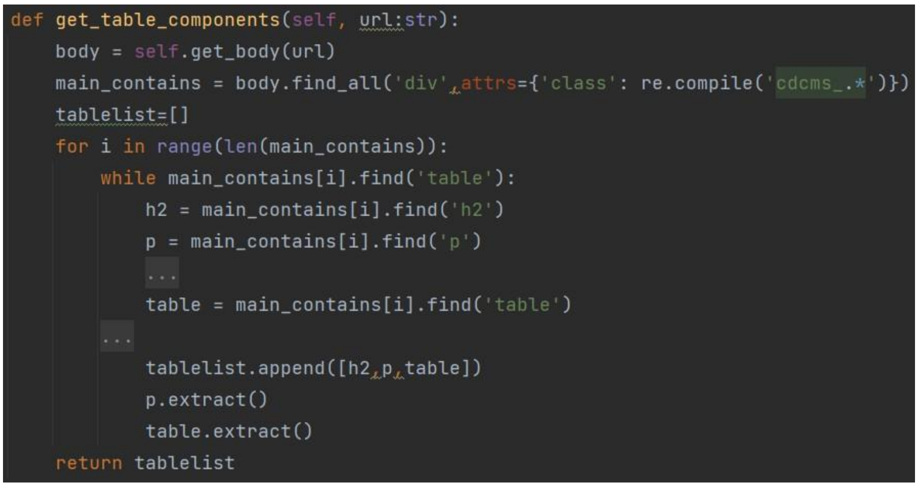
```
['<h1>', 'UCLA', '</h1>', '<h2>', 'UCI', '</h2>', '<h3>', 'UCB', '</h3>', '<p>',
'The', 'University', 'of', 'California', 'is', 'the', 'most', 'influential', 'public',
'university', 'system', 'in', 'the', 'world.', '</p>', '<h2>', 'UCB', '</h2>', '<h2>',
'UCB', '</h2>']
```

```
[0,1,1,2,3,3,4,5,5,6,7,8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,9,10,10,11,12,12,13]
```

Figure 13. The o

### Retrieve Table

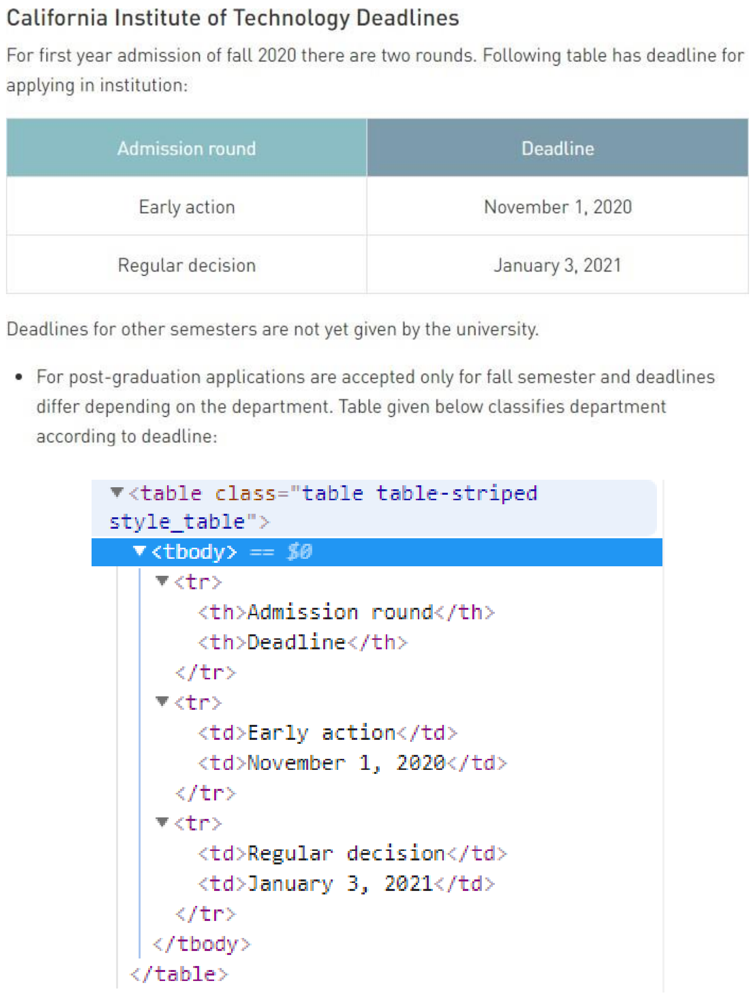
In order to get some tables on the page, we first used get_ table_ components method, then used find ('table ') to put all the tables in the page into a list.



```
def get_table_components(self, url:str):
    body = self.get_body(url)
    main_contains = body.find_all('div',attrs={'class': re.compile('cdcms_.*')})
    tablelist=[]
    for i in range(len(main_contains)):
        while main_contains[i].find('table'):
            h2 = main_contains[i].find('h2')
            p = main_contains[i].find('p')
            ...
            table = main_contains[i].find('table')
        ...
            tablelist.append([h2,p,table])
            p.extract()
            table.extract()
    return tablelist
```

Figure 14. The o

Because most of the tables are made with HTML tags such as < tr > < td >, we chose to save these tables by row. we first use find ('tr') to determine the rows of the table, and then use find ('td') or find ('th') to take out all the elements of each row and put them into a list. Finally, we put all the lists representing rows into a list representing table, and we get a two dimensional list containing all the information of this table, because values in the list can be obtained through index, it is very convenient for me to read the information in these tables.



California Institute of Technology Deadlines

For first year admission of fall 2020 there are two rounds. Following table has deadline for applying in institution:

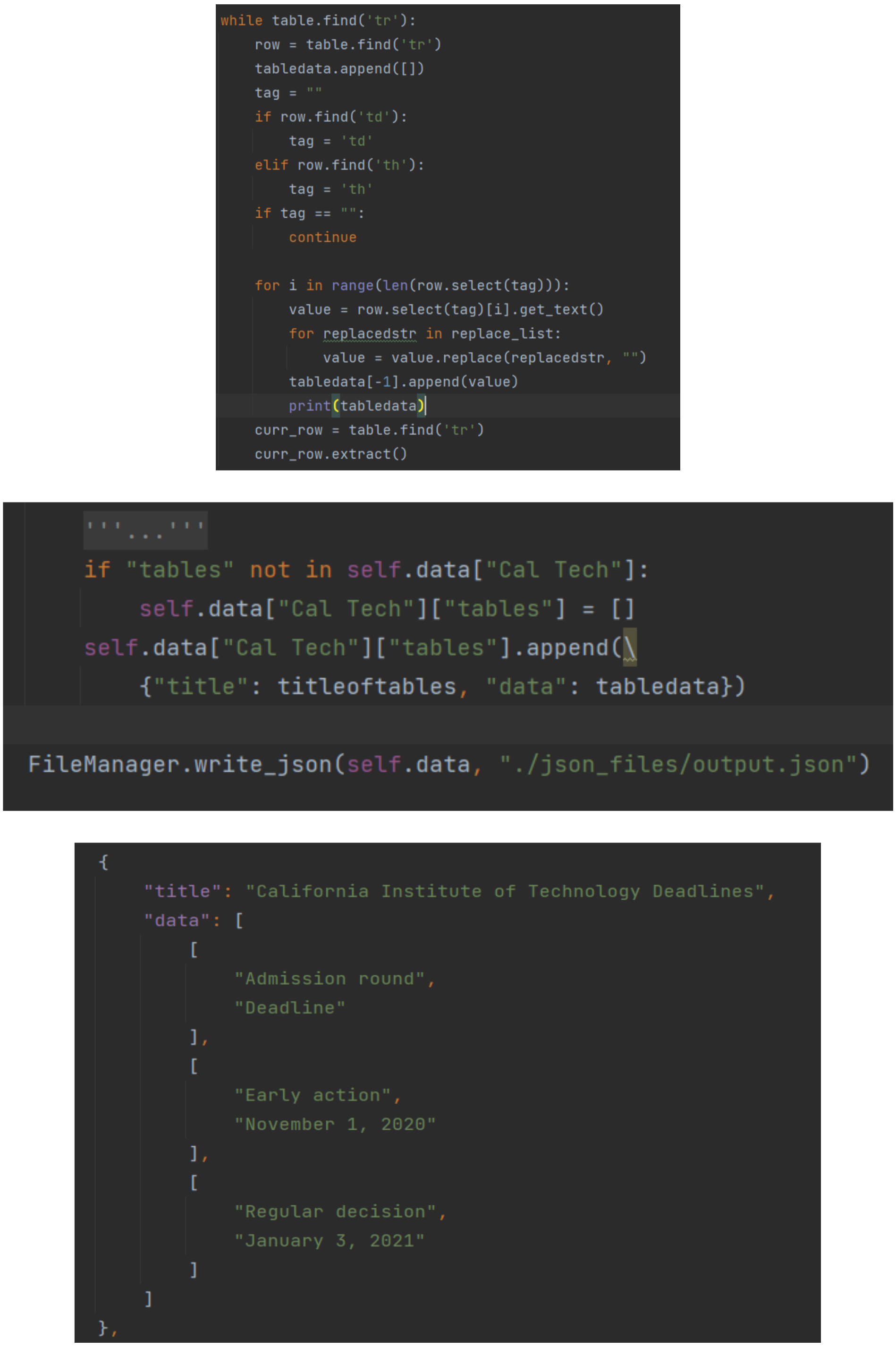
| Admission round | Deadline |
|---|---|
| Early action | November 1, 2020 |
| Regular decision | January 3, 2021 |

Deadlines for other semesters are not yet given by the university.

- For post-graduation applications are accepted only for fall semester and deadlines differ depending on the department. Table given below classifies department according to deadline:

```
<table class="table table-striped
style_table">
  <tbody> == $0
    <tr>
      <th>Admission round</th>
      <th>Deadline</th>
    </tr>
    <tr>
      <td>Early action</td>
      <td>November 1, 2020</td>
    </tr>
    <tr>
      <td>Regular decision</td>
      <td>January 3, 2021</td>
    </tr>
  </tbody>
</table>
```

Figure 15. The code excerpt for the forntend web pages



```
while table.find('tr'):
    row = table.find('tr')
    tabledata.append([])
    tag = ""
    if row.find('td'):
        tag = 'td'
    elif row.find('th'):
        tag = 'th'
    if tag == "":
        continue

    for i in range(len(row.select(tag))):
        value = row.select(tag)[i].get_text()
        for replacedstr in replace_list:
            value = value.replace(replacedstr, "")
        tabledata[-1].append(value)
        print(tabledata)
    curr_row = table.find('tr')
    curr_row.extract()
```

```
    '''...'''
    if "tables" not in self.data["Cal Tech"]:
        self.data["Cal Tech"]["tables"] = []
    self.data["Cal Tech"]["tables"].append(\
        {"title": titleoftables, "data": tabledata})

FileManager.write_json(self.data, "./json_files/output.json")
```

```
{
    "title": "California Institute of Technology Deadlines",
    "data": [
        [
            "Admission round",
            "Deadline"
        ],
        [
            "Early action",
            "November 1, 2020"
        ],
        [
            "Regular decision",
            "January 3, 2021"
        ]
    ]
},
```

Figure 16. The code excerpt for JSON data

## 4. Related Work

Stella and Woodhouse discussed the issues related with public ranking for higher education institutions [14]. They pointed out that the ranking may not effectively reflect the true quality of the institutions, which aligns the purpose of our work. However, they did not use any big data to draw the conclusion. Our approach focuses on using real school data and automated framework to get a more objective result.

The researchers Aguillo and Orduna-Malea discussed how they make a development of the "Would Class University and The ranking web"[15]. In their research paper, they indicate that the web application can assess most of the top universities by using G-factor which is a statistical factor that captures the diversity of motivations. However, unlike doing web crawling and noise removal in our application, their application doesn't do open source updates from the ranking universities.

Gupta and Divakar provided an imported approach to ranking web documents in 2012. Their ranking improved from their excremental data from the overall search results [16]. The ordering process will utilize the previous ranking score and train the new data with the previous data to make the new ranking more credible. However, their ranking is mostly finished with HTML only, but no framework like flask that our application used to make the web app more manageable and flexible.

## 5. Conclusions And Future Work

The purpose of DSP application is to create an information integration center for students to reduce the complexity of applying to undergraduate or graduate university. In the web application, we included top rank university information for admission, so that students don't need to search by themselves to collect admission requirements.

To retrieve information from the official university website, we applied Web crawling to scrap the amount of weblinks. During the Web crawling process, we used BFS with a data structure queue to traverse web outlinks. Beautiful soup, a tool based on pythons, can help me to get all outlinks from root links and do basic parsing. we utilized the algorithm of noise removal to solve the problem of obtaining all the main contents and eliminating the useless information, such as ads, menu bar, header, footer etc. In order to make my application understandable and manageable, we used the flask framework to build the skeleton of the applications. In addition, we used Python as the main development language, because all the techniques we am using, like beautiful soup and flask, fit with Python. HTML, CSS, and JavaScript is used for the frontend of my web application. To create tables to integrate the information that we retrieved from information retrieval [3]. we used jQuery to make the creating table function reusable with all various tables and render that with HTML.

My application is deployed with Heroku.com. A couple of users and my friends viewed my website, and the feedback is good so far. The next step for developing the application is to create a community for users to char or share their experience. The final goal of the applications should be a community with full functionality for university admission.

Although this tool can effectively extract information for most of the application information pages composed of text and simple forms, there are still some pages made of pictures or more complex forms, and the extraction of this part of information may not be accurate and complete.

The function of this tool is relatively simple. If applicants can select institutions according to their scores and background, the application efficiency will be further improved.

Next, we hope to continue to develop new functions for this tool, such as the function of screening and matching schools, and improve the existing code to improve the applicability and ensure that all forms of information can be crawled. In addition, we will improve my algorithm to further speed up the crawling progress, so that this tool can maximize its advantages.

## REFERENCES

[1] M. Grinberg, Flask web development: Developing advanced web applications with python. Sebastopol, CA: O'Reilly Media, 2014.

[2] W. B. Croft, D. Metzler, and T. Strohman, Search engines: Information retrieval in practice. Addison-Wesley Professional, 2011.

[3] "Information Retrieval Systems," Sciencedirect.com. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/information-retrieval-systems. [Accessed: 23-Jul-2021].

[4] N. Nurseitov, M. Paulson, R. Reynolds, and C. Izurieta, "Comparison of JSON and XML data interchange formats: A case study," Montana.edu. [Online]. Available: https://www.cs.montana.edu/izurieta/pubs/caine2009.pdf. [Accessed: 23-Jul-2021].

[5] H. K. Azad, R. Raj, R. Kumar, H. Ranjan, K. Abhishek, and M. P. Singh, "Removal of noisy information in web pages," in Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies - ICTCS '14, 2014.

[6] W. Cui, L. Huang, L. Liang, and J. Li, "The research of PHP development framework based on MVC pattern," in 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, 2009, pp. 947–949.

[7] "Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation," Crummy.com. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed: 23-Jul-2021].

[8] "Lxml - processing XML and HTML with Python," Lxml.de. [Online]. Available: https://lxml.de/. [Accessed: 23-Jul-2021].

[9] "How Google's site crawlers index your site - Google search," Google.com. [Online]. Available: https://www.google.com/search/howsearchworks/crawling-indexing/. [Accessed: 23-Jul-2021].

[10] S. Beamer, K. Asanovic, and D. Patterson, "Direction-optimizing breadth-first search," in 2012 International Conference for High Performance Computing, Networking, Storage and Analysis, 2012, pp. 1–10.

[11] G. L. Hajba, Website scraping with python: Using BeautifulSoup and scrapy, 1st ed. Berlin, Germany: APress, 2018.

[12] "FILO," Techterms.com. [Online]. Available: https://techterms.com/definition/filo. [Accessed: 23-Jul- 2021].

[13] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," arXiv [cs.CL], 2002.

[14] "Ranking of higher education institutions / Antony Stella and David Woodhouse," Gov.au. [Online]. Available: https://catalogue.nla.gov.au/Record/5784960. [Accessed: 22-Jul-2021].

[15] I. F. Aguillo and E. Orduña-Malea, "The ranking web and the 'world-class' universities," in Building World-Class Universities, Rotterdam: SensePublishers, 2013, pp. 197–217.

[16] P. Gupta, S. K. Singh, D. Yadav, and A. K. Sharma, "An improved approach to ranking web documents," J. Inf. Process. Syst., vol. 9, no. 2, pp. 217–236, 2013.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# MASK REGION-BASED CONVOLUTIONAL NEURAL NETWORKS (R-CNN) FOR SINHALA SIGN LANGUAGE TO TEXT CONVERSION

R. D. Rusiru Sewwantha[1] and T. N. D. S. Ginige[2]

[1]Universal College Lanka, University of Central Lancashire,
Colombo, Sri Lanka
[2]Universal College Lanka, Colombo, Sri Lanka

## *ABSTRACT*

*Sign Language is the use of various gestures and symbols for communication. It is mainly used by disabled people with communication difficulties due to their speech or hearing impediments. Due to the lack of knowledge on sign language, natural language speakers like us, are not able to communicate with such people. As a result, a communication gap is created between sign language users and natural language speakers. It should also be noted that sign language differs from country to country. With American sign language being the most commonly used, in Sri Lanka, we use Sri Lankan/Sinhala sign language. In this research, the authors propose a mobile solution using a Region Based Convolutional Neural Network for object detection to reduce the communication gap between the sign users and language speakers by identifying and interpreting Sinhala sign language to Sinhala text using Natural Language Processing (NLP). The system is able to identify and interpret still gesture signs in real-time using the trained model. The proposed solution uses object detection for the identification of the signs.*

## *KEYWORDS*

*Object detection, TensorFlow, Mask R-CNN, Sinhala, Sign Language, VideoCapture.*

## 1. INTRODUCTION

Essentially, the use of gestures and symbols for communication is known as sign language. World health organization has revealed that over 5% of the world population (430 million people) require rehabilitation to address their ‘disabling’ hearing loss, it was also mentioned that by the year 2050, over 700 million people will have a disabling hearing loss [1]. In 2020, it was stated by the World Federation of the Deaf that there are around 72 million people worldwide who uses sign language for communication [2]. According to the World health organization, approximately 9% of the Sri Lankan population has loss of hearing [3] and from a study conducted by the Sri Lanka Federation of the Deaf, more than 300,000 people are deaf. Unfortunately, there are not many sign language interpreters in Sri Lanka. It was mentioned that there were only 6 sign interpreters in Sri Lanka [4]. Due to this, there is a huge demand for sign interpreters in our country. Even though there have been many proposed systems to tackle this problem, a proper system has not been deployed yet. Therefore, the goal of this project is to develop a mobile Sinhala sign language interpreter which can be easily used by language speakers or disabled people to communicate.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 207-219, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111418

The objective and the core functionality of the system is that it should be able to successfully interpret inputted Sinhala signs into Sinhala text. To cater this requirement, the following features are implemented in the prototype.

- Real-time sign interpretation using video capture and video upload from gallery.
- Sign language dictionary.
- Trained mask rcnn inception resnet model for object detection.

The purpose of the sign dictionary is to make sure that the language speakers are able to learn the basic signs of Sinhala sign language and to inform users on what signs the system can interpret in the current version.

## 2. LITERATURE REVIEW

Sign language interpretation can be identified as gesture recognition. Since gesture recognition has been an area of interest for many years, several methods have been employed [5]. These methods can be categorized into two as data glove method and vision-based method [5].

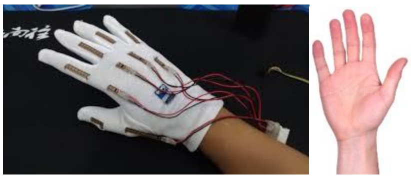


Figure 1. Data glove method & Vision based method [6].

### 2.1. Data Glove Methods

Data glove method employs mechanical or optical sensors attached to a glove that transforms finger flexions into electrical signals to determine the hand posture [6]. This approach is not a mobile solution.

Using the data glove method, a Chinese sign language recognition system was built based on ARM9 [7]. It also uses combined flex sensors with 9-axis IMU sensor. The sensor modules can measure both the bending degree of the fingers and the angle of the palm. These readings are sent to the ARM9 processor which will analyse and process the data in real-time. Finally, using a voice broadcast module and the text display module, the interpreted text and voice are displayed and broadcasted.

A data glove with gyro-sensor was introduced for Japanese sign language [8]. The glove was developed to have 5 sensors for the fingers and on the back of the palm, an accelerometer with gyro-sensor, was installed. These data are sent through an analog-to-digital converter and then to the data control unit. From this the data is sent to a computer via a USB cable. This also consists of a palm turning motion algorithm.

### 2.2. Vision-Based Methods

Comparatively, vision-based approach uses cameras to input images [6]. The gestures used to create the gesture database should be selected with their relevant meaning and each gesture may contain multiple samples for increasing the accuracy of the system.

Nath & Arun, 2017 proposes a sign language interpreter implemented in ARM CORTEX A8 processor board using convex hull algorithm and template matching algorithm. A webcam is used to feed images into the system and these images are converted into text. Numbers are recognized by the convex hull method and alphabets are recognized by the template matching method. This system consists of several physical components (camera, display device, processor, etc.).

An interpreter for American sign language developed by Ahmed et. al. (2016) using Microsoft’s Kinect V2's Continuous Gesture Builder for gesture recognition and training has two modules, speech to sign and sign to speech. Speech to sign is done using an external library for speech to text conversion by taking a sentence or a word as input. Keywords are identified from the input and compared against keywords in the database to identify the gesture against the keyword. These gestures are mapped with the animations to a 3d model using a data structure in Unity3D. For the sign to speech module, the gestures are inputted frame by frame using the Kinect sensor and matched it with the pre-stored gestures in the database. By using these keywords, a sentence is constructed and is converted into speech by text to speech libraries provided by .Net.

A Sinhala sign language framework is proposed by Madushanka et. al. (2016) using a wearable armband. This research is an extension of Surface Electromyography (sEMG) - used to gather gestural data and Inertial Measurement units (IMU) - (Accelerometer gyroscope and Magnetometer) are used for spatial data. For this system, a device called “Myo Gesture Recognition Armband” (developed by Thalmic Labs Inc.) has been used. This device is Bluetooth compatible and uses sEMG as the main gesture recognition technology. It also has a combination of sEMG, accelerometer, gyroscope, and magnetometer sensors. For gestural references, sEMG data has been used and for spatial references accelerometer, gyroscope and magnetometer has been used. For gesture recognition, a supervised machine learning technique has been used.

Dissanayake et. al. (2020) proposed a mobile app for Sinhala sign language interpretation using image processing and machine learning called “Utalk”. This solution converts signs from videos to Sinhala text. This also can interpret both static and dynamic signs. A signing video is taken as the input to the system and then the frame segments are extracted from it so that the using image processing techniques, the background of these frames can be removed. Then these pre-processed images are fed into two separate machine learning models as static sign classifier and dynamic sign classifier by classifying them as static and dynamic signs. These are fed into the language model which is used to generate text based on the input video.

The above mentioned are the previous work conducted on sign language interpretation. Most of them have been proposed for American sign language and almost all solutions contain hardware components. Having hardware components are cumbersome, resulting in inability to use the products in the day-to-day life. However, the proposed solution by the authors of this research, is a mobile application. Therefore, it can be easily used by anyone with a smartphone. It is also based on Sinhala sign language which is the sign language used in Sri Lanka. Currently in Sri Lanka, a proper usable solution has not been implemented.

## 3. METHODOLOGY

### 3.1. Data Gathering

The image dataset prepared is the most important requirement in a data science project. No proper datasets were found for the Sinhala sign language to download from internet. Therefore, images were collected from several people posing the signs required to train the model. When capturing the images, they were advised to take each sign from several angles. This is to make sure that the model is trained to identify signs from different angles. Following tables display which signs were used to train the model.

Table 1. Number Signs

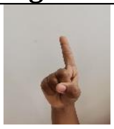
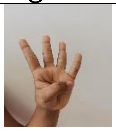
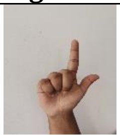
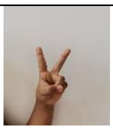
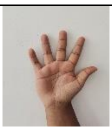
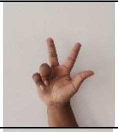
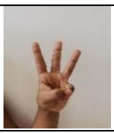
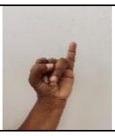
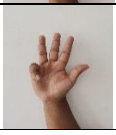
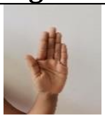
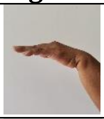
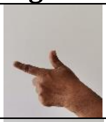
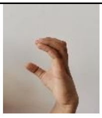
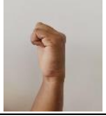
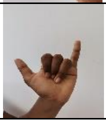
| Value | Sign | Value | Sign | Value | Sign |
|---|---|---|---|---|---|
| 1 |  | 4 |  | 7 |  |
| 2 |  | 5 |  | 8 |  |
| 3 |  | 6 |  | 9 |  |

Table 2. Alphabet Signs

| Value | Sign | Value | Sign | Value | Sign |
|---|---|---|---|---|---|
| අ |  | ඉ |  | ට |  |
| ආ |  | එ |  | ශ |  |

## 3.2. Technology Selection

### 3.2.1. Development Framework

From the choice between Django and Flask frameworks, Flask was selected to develop the backend Rest API for the sign interpretation system. Reasons for the selection of Flask framework instead of Django was due to its lightweight property compared to Django. Furthermore, using Flask to create HTTP services was easier. Also, for a Python beginner, Flask framework is easier to learn than Django which is rich in features.

Since the frontend is a mobile application, to make the code reusable, Ionic Framework was decided to be used as it is a cross-platform framework. Main reason for selecting this was that, using the same implementation, the application can be easily deployed on both Android and iOS or as a web application. Another reason to select this was that the authors had prior knowledge on Ionic framework.

### 3.2.2. Machine Learning Library

As the machine learning library, Tensorflow 2 was selected. This was used to implement the object detection of the system, which is the core functionality of the sign interpretation system, which is the sign identification. It is also possible to use the Keras library with the Tensorflow library.

## 3.3. Pre-Processing

After the images were collected for the dataset, identical images were removed to reduce the overfitting of the model when training. Then all the images were resized to the dimensions 900x1024. Afterwards, in the ratio 0.2 to 0.8, the dataset was divided into two as test and train, respectively. These divided images were annotated to train the mask R-CNN model, which will also increase the accuracy of the training process, outlining the signs. Each sign image was annotated and labelled as its sign text value using the annotation tool LabelMe. Figure 8 shows how this was done.

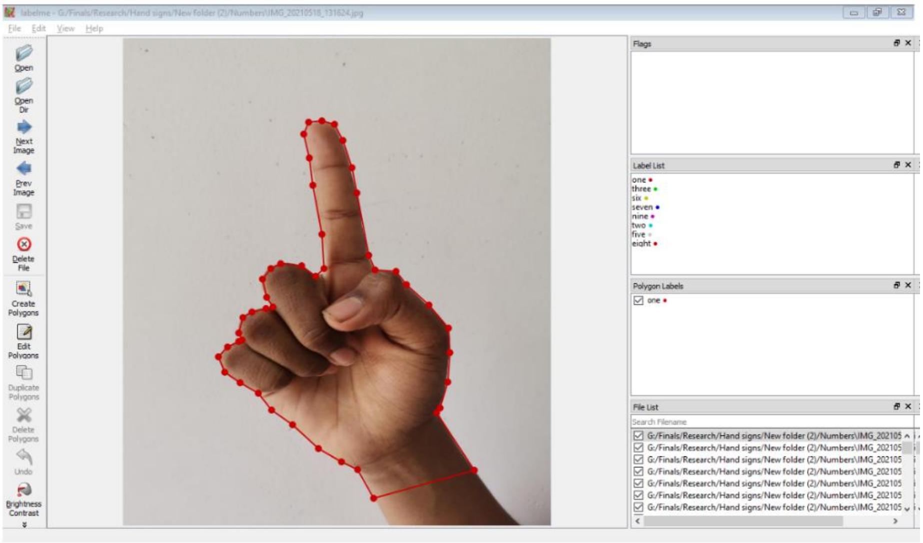

labelme - G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_20210518_131624.jpg
File
Edit
View
Help
Open
Open Dir
Next Image
Prev Image
Save
Delete File
Create Polygons
Edit Polygons
Duplicate Polygons
Delete Polygons
Undo
Brightness Contrast
Flags
Label List
one ●
three ●
six ●
seven ●
nine ●
two ●
five ●
eight ●
Polygon Labels
☑ one ●
File List
Search Filename
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105
☑ G:/Finals/Research/Hand signs/New folder (2)/Numbers\IMG_202105

Figure 2. Label Me annotation tool

After saving the annotated image as a json file, all the images were converted to Common Object in Context (COCO) format. Finally, this dataset is converted to TFRecords, Tensorflow's binary storage format, which increases the performance and training speed of the model.

## 3.4. Implementing the Model

To train the custom model for the sign identification, a pretrained model which is best for this project was selected from the Tensorflow model zoo. This is because a pre-trained model is previously trained using a large-scale image classification task. To serve the purpose of Sinhala sign detection, transfer learning was used which is used to customize these pretrained models. The pretrained model selected for this purpose was mask_rcnn_inception_resnet_v2 model, with the highest COCO mean average precision.

The configuration file, mask_rcnn_inception_resnet_v2_1024x1024_coco17_gpu-8, for training the model was available in the Tensorflow 2 GitHub repository in the 'configs' directory. This file was updated to meet the requirements needed to train the model for this project. Finally, Google Colaboratory was used to train the model with a GPU accelerated compute engine. Figure 7 shows the architecture of a mask R-CNN model. This trained model is loaded into the backend (Python API) of the system since it requires a considerable amount of processing power.

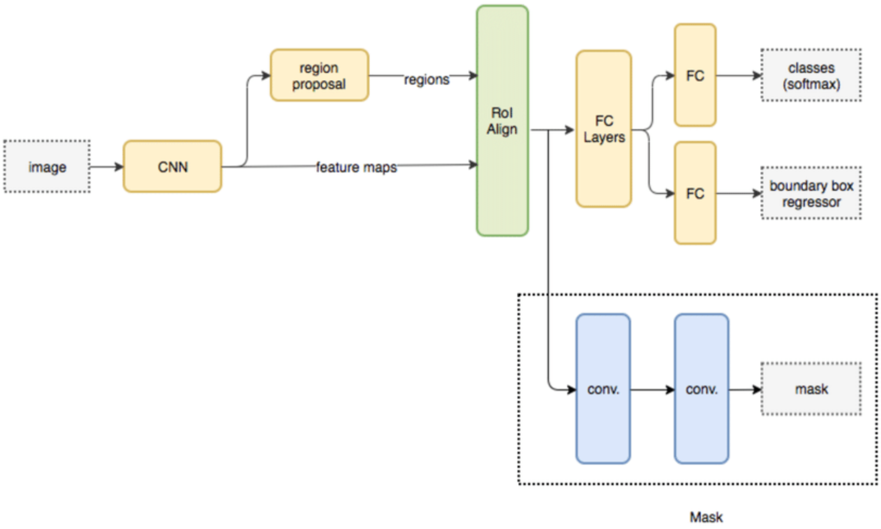


Figure 3. Architecture of the mask R-CNN framework [4]

## 3.5. Implementation of the Interpretation

The frontend of the system, mobile application, is used to feed data into the system. Interpretation occurs on the backend API of the system.

OpenCV VideoCapture was used to capture the video feed from the user. Since the API and database was not hosted on a Cloud Server, this was assigned to use the camera from the machine that the API is running since the server should have high GPU performance for the functionality of this feature.

Once the video is captured into the system, it will be separated into frames and processed to detect the sign in each frame by sending the frames through the loaded model. Later, the text value of the detected sign is retrieved from the database using the label class of the detected sign. Figure 4 shows the flow diagram for the interpretation process in the system.

Another feature of the system is the sign dictionary. For this a GET request is sent from the frontend to the API. After receiving the request, the API will execute the query to return all the data available in the database. This will display a list of signs and their values.

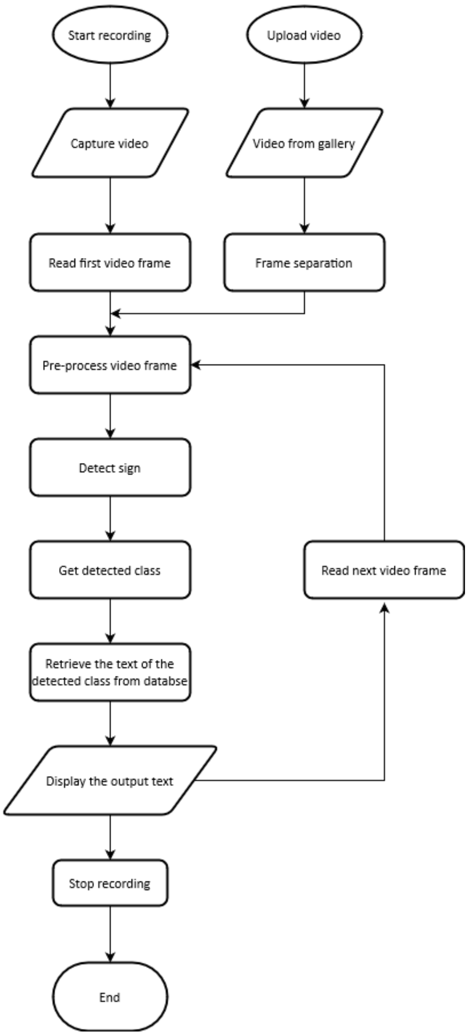


Figure 4. Process flow of the interpretation

## 3.6. Final Prototype

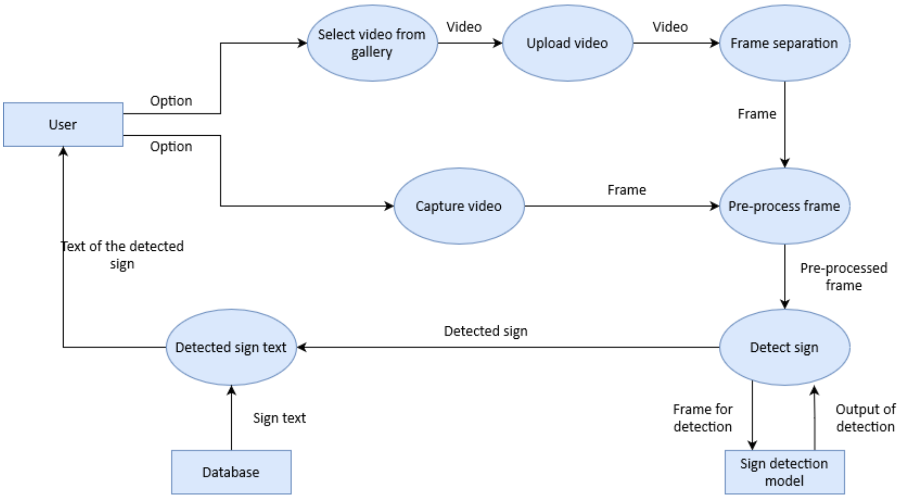


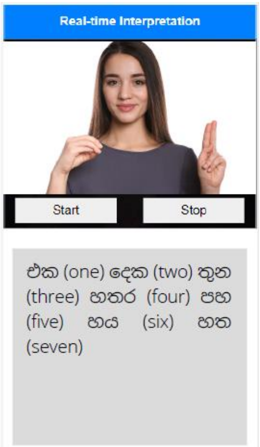
Figure 5. 2-Level Data Flow Diagram

Figure 5 shows the 2-Level Data flow diagram of the system that explains the flow of the core functionality of the system, which is real-time interpretation of Sinhala sign language, by showing both the core and sub functionalities.

In this diagram, the user is considered as an external entity. User is given two options, to upload video from the gallery or capture real-time video and upload to the system, for the interpretation to happen. If a video is uploaded from the gallery, it is separated into frames and pre-processed. If the user selects to capture real-time video, since the video is captured as frames, the frames are sent straight to the pre-processing stage. Once the frames are processed, they are sent to the detection model for detection. Detected sign value is then retrieved and using the database, the text value of the detected sign is retrieved and displayed to the user.

### 3.6.1. User Interface

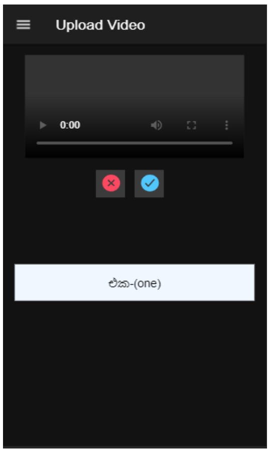
User interface was designed to be simple and easy to use. Below are the screenshots of the UI of the final prototype.



එක (one) දෙක (two) තුන (three) හතර (four) පහ (five) හය (six) හත (seven)

Figure 6. Real-time interpretation UI

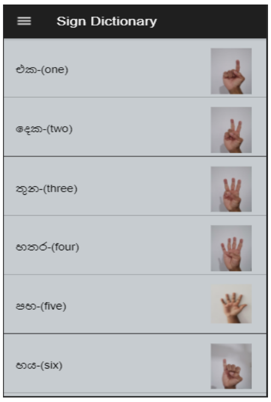
Figure 6 shows the real-time interpretation UI after performing an interpretation.



එක-(one)

Figure 7. Video interpretation UI

Figure 7 shows the UI and an example output of the video interpretation.



Sign Dictionary

එක-(one)

දෙක-(two)

තුන-(three)

හතර-(four)

පහ-(five)

හය-(six)

Figure 8. Sign dictionary UI

Figure 8 shows the list of signs with its UI in the sign dictionary.

# 4. EVALUATION

## 4.1. Evaluation Data and Process

### 4.1.1. Accuracy

The accuracy was tested for each sign that the model was trained. This was done for both the real-time and video interpretation features. For the real-time interpretation, each sign was tested with 10 different users where each user performed the same sign for 5 times. For the video interpretation, the same 10 users were used to record 10 small video clips with each person signing. Accuracy values were taken as the average calculated from the results generated from the mentioned process. Below table shows the calculated average.

Table 3. Accuracy readings for real-time sign interpretation

| Sign used | Percentage accuracy |
|---|---|
| 1 | 71% |
| 2 | 75% |
| 3 | 73% |
| 4 | 75% |
| 5 | 82% |
| 6 | 79% |
| 7 | 83% |
| 8 | 76% |
| 9 | 88% |
| අ | 70% |
| ආ | 71% |
| ඉ | 84% |
| එ | 79% |
| උ | 76% |
| ය් | 85% |

Overall accuracy of the trained model for real-time interpretation – 77.8%

### 4.1.2. Usability

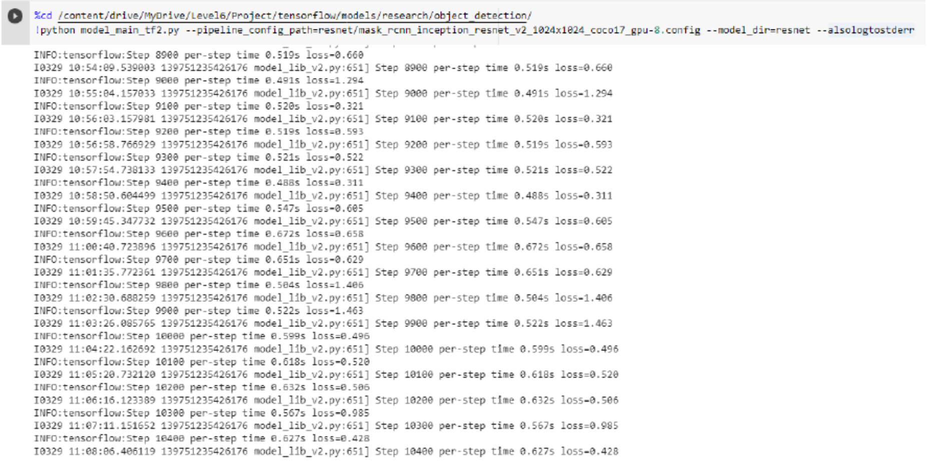
To test the usability and responsiveness of the mobile application, it was installed and tested in several android mobile devices. The user interface was displayed without any issues and was responsive irrespective of the screen resolution. Same process was done by running the application as a web application. This was also successful without any problems. The flow of the application was working without any dead ends. Since this application will be mostly used by language speakers, for testing random people were selected who are natural language speakers. Below is the feedback received from the users on usability.

"*The application that was developed is surprisingly easy to learn and use." – Mr. Thathsara Nandun.*

*"The user interface of the system can be understood easily and without any prior knowledge of the system, it could be easily used." – Mrs. Chandra Jayanthi.*

## 4.2. Evaluation of Machine Learning Model

The Jupyter notebook created was loaded from the Google Colaboratory. The required libraries and protocol buffers were installed after cloning the Tensorflow GitHub repository to the google drive and mounting the drive to the Google Colaboratory. The training of the model was stopped when the graph curve started to flatten, making small changes in the loss value at 10460 steps with each step being batch size of 1 feeding the model with an image per step and until the loss reading was less than zero. This was done to prevent the model from overfitting. Finally, the model was exported for inference. Figure 9 shows the output generated during the model training.



```
%cd /content/drive/MyDrive/Level6/Project/tensorflow/models/research/object_detection/
!python model_main_tf2.py --pipeline_config_path=resnet/mask_rcnn_inception_resnet_v2_1024x1024_coco17_gpu-8.config --model_dir=resnet --alsologtostderr
INFO:tensorflow:Step 8900 per-step time 0.519s loss=0.660
I0329 10:54:09.539003 139751235426176 model_lib_v2.py:651] Step 8900 per-step time 0.519s loss=0.660
INFO:tensorflow:Step 9000 per-step time 0.491s loss=1.294
I0329 10:55:04.157033 139751235426176 model_lib_v2.py:651] Step 9000 per-step time 0.491s loss=1.294
INFO:tensorflow:Step 9100 per-step time 0.520s loss=0.321
I0329 10:56:03.157981 139751235426176 model_lib_v2.py:651] Step 9100 per-step time 0.520s loss=0.321
INFO:tensorflow:Step 9200 per-step time 0.519s loss=0.593
I0329 10:56:58.766929 139751235426176 model_lib_v2.py:651] Step 9200 per-step time 0.519s loss=0.593
INFO:tensorflow:Step 9300 per-step time 0.521s loss=0.522
I0329 10:57:54.738133 139751235426176 model_lib_v2.py:651] Step 9300 per-step time 0.521s loss=0.522
INFO:tensorflow:Step 9400 per-step time 0.488s loss=0.311
I0329 10:58:50.604499 139751235426176 model_lib_v2.py:651] Step 9400 per-step time 0.488s loss=0.311
INFO:tensorflow:Step 9500 per-step time 0.547s loss=0.605
I0329 10:59:45.347732 139751235426176 model_lib_v2.py:651] Step 9500 per-step time 0.547s loss=0.605
INFO:tensorflow:Step 9600 per-step time 0.672s loss=0.658
I0329 11:00:40.723896 139751235426176 model_lib_v2.py:651] Step 9600 per-step time 0.672s loss=0.658
INFO:tensorflow:Step 9700 per-step time 0.651s loss=0.629
I0329 11:01:35.772361 139751235426176 model_lib_v2.py:651] Step 9700 per-step time 0.651s loss=0.629
INFO:tensorflow:Step 9800 per-step time 0.504s loss=1.406
I0329 11:02:30.688259 139751235426176 model_lib_v2.py:651] Step 9800 per-step time 0.504s loss=1.406
INFO:tensorflow:Step 9900 per-step time 0.522s loss=1.463
I0329 11:03:26.085765 139751235426176 model_lib_v2.py:651] Step 9900 per-step time 0.522s loss=1.463
INFO:tensorflow:Step 10000 per-step time 0.599s loss=0.496
I0329 11:04:22.162692 139751235426176 model_lib_v2.py:651] Step 10000 per-step time 0.599s loss=0.496
INFO:tensorflow:Step 10100 per-step time 0.618s loss=0.520
I0329 11:05:20.732120 139751235426176 model_lib_v2.py:651] Step 10100 per-step time 0.618s loss=0.520
INFO:tensorflow:Step 10200 per-step time 0.632s loss=0.506
I0329 11:06:16.123389 139751235426176 model_lib_v2.py:651] Step 10200 per-step time 0.632s loss=0.506
INFO:tensorflow:Step 10300 per-step time 0.567s loss=0.985
I0329 11:07:11.151652 139751235426176 model_lib_v2.py:651] Step 10300 per-step time 0.567s loss=0.985
INFO:tensorflow:Step 10400 per-step time 0.627s loss=0.428
I0329 11:08:06.406119 139751235426176 model_lib_v2.py:651] Step 10400 per-step time 0.627s loss=0.428
```

Figure 9. Model training output

## 4.3. Evaluation of the Final System

The research aim was to implement a Sinhala Sign Language interpreter by designing, developing, and evaluating Artificial Intelligence models and finally use them in building a mobile which will be able to successfully identify and interpret Sinhala signs in real-time.

The following features were implemented for this prototype of the system.

- Real-time sign interpretation using video capture and video upload from gallery.
- Sign language dictionary.
- Trained mask rcnn inception resnet model for object detection.

In the implementation of the prototype, few limitations were identified. They are,

- The accuracy of the sign identification depends on the light conditions of the captured/uploaded video.
- The video capture feature is working as a web application, and this feature was not deployed on a mobile device.
- Due to the lack of resources such as a GPU, the speed of the interpretations was longer than expected.
- Sign identification only works for still gesture signs.

After identifying the limitations of the system, the following future enhancements are discussed.

- Improve the accuracy of the system based on the light factor of the input video.
- Implement and deploy a mobile version for the video capturing feature.
- Increase the speed of the interpretations by hosting the API on a cloud server.
- Implement and train the model for identifying dynamic signs.

This was how the research was done, a prototype was implemented and tested, its limitations were identified, and future enhancements were discussed.

## 5. CONCLUSION

The research aim was to implement a Sinhala Sign Language interpreter by designing, developing, and evaluating Artificial Intelligence models and finally use them in building a mobile or web application which will be able to successfully identify and interpret Sinhala signs in real-time. This was successfully achieved by implementing the system.

Implementing this research project has also benefitted the authors by helping to learn new tools and technologies and improve existing skills. Knowledge needed was gained by referring the documentations of each technology and testing the knowledge gained by implementing simple functionalities. Usage of Tensorflow library was studied deeply for the purpose of this research.

During the process, one of the challenges faced was the lack of datasets to train the model. As a result, a custom dataset was developed by the authors. Another problem was the lack of resources to train the model. To overcome one such challenge, Google Colaboratory was used to train the model without any issues.

The authors hope this will be a good solution to identify and interpret signs from Sinhala sign language and by doing so, helping to reduce the communication gap between sign users and language speakers.

## ACKNOWLEDGEMENTS

We would like to extend our gratitude towards our parent university, University of Central Lancashire for providing us with the opportunity and resources to complete this research successfully. Our appreciation goes to the staff of Universal College Lanka for supporting and advising us whenever needed.

We would also like to thank our colleagues for supporting and encouraging us at times required and the candidates that volunteered to collect the data needed and support when testing the system. Last but not least, we would like to thank our parents for their continuous support and the motivation given to us throughout the project.

## REFERENCES

[1] World health organization. (2021, April 1). Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] Ahmed, M., Idrees, M., ul Abideen, Z., Mumtaz, R., Khalique, S. (2016). Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's kinect v2. IEEE. http://dx.doi.org/10.1109/SAI.2016.7556002.

[3] Sinhala Sign Language the main communication mode for the Deaf in Sri Lanka. (2019, January 18). DailyFT. http://www.ft.lk/opinion/Sinhala-Sign-Language-the-main-communication-mode-for-the-Deaf-in-Sri-Lanka/14-671078.

[4] Rupasinghe, H. (2018, March 27). Sri Lanka Terribly short of sign language interpreters. *Dailymirror*.

[5] Nath, G. G., Arun, C. S. (2017). Real time sign language interpreter. IEEE. http://dx.doi.org/10.1109/ICEICE.2017.8191869.

[6] Ibraheem, N. A., Khan, R. Z. (2021). Vision Based Gesture Recognition Using Neural Networks Approaches*: A Review*. https://www.researchgate.net/profile/Noor-Ibraheem/publication/267991106.

[7] Lei, L. Dashun, Q. Design of data-glove and Chinese sign language recognition system based on ARM9. 2015 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 2015, pp. 1130-1134, http://dx.doi.org/10.1109/ICEMI.2015.7494440.

[8] Mori, Y. Toyonaga, M. Data-Glove for Japanese Sign Language Training System with Gyro-Sensor. 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), 2018, pp. 1354-1357, http://dx.doi.org/10.1109/SCIS-ISIS.2018.00211.

[9] Nath, G. G., Arun, C. S. (2017). Real time sign language interpreter. IEEE. http://dx.doi.org/10.1109/ICEICE.2017.8191869.

[10] Ahmed, M., Idrees, M., ul Abideen, Z., Mumtaz, R., Khalique, S. (2016). *Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's kinect v2*. IEEE. http://dx.doi.org/10.1109/SAI.2016.7556002.

[11] Madushanka, A. L. P., Senevirathne, R. G. D. C., Wijesekara, L. M. H., Arunatilake, S. M. K. D., Sandaruwan, K. D. (2016). *Framework for Sinhala Sign Language recognition and translation using a wearable armband*. IEEE. http://dx.doi.org/10.1109/ICTER.2016.7829898.

[12] Dissanayake, I. S. M., Wickramanayake, P. J., Mudunkotuwa, M. A. S., Fernando, P. W. N. (2020). *Utalk: Sri Lankan Sign Language Converter Mobile App using Image Processing and Machine Learning.* IEEE. http://dx.doi.org/10.1109/ICAC51239.2020.9357300.

## AUTHORS

**Rusiru Sewwantha** is currently a BSc (Hons) in Software Engineering 4th year Undergraduate at Universal College Lanka affiliated with University of Central Lancashire, UK. His passion towards programming and machine learning has made it possible for him to complete this research.

**Thepul Ginige** is currently reading for his Ph.D. (University of Colombo), has a Master of Information Technology (University of Colombo), Pg. Dip in Criminology & Criminal Justice (University of Sri Jayewardenepura), Postgraduate diploma national university of Singapore, Bachelor of Science (Hons) (Information Technology Sri Lanka Institute of Information Technology), Bachelor of Science (University of Sri Jayewardenepura). He is currently working as a Senior Lecturer / Programme Coordinator– Universal College Lanka. Former -Senior Lecturer / Manager Special Project -Saegis Campus, Senior Lecturer / Academic Coordinator (Scottish Qualifications Authority -SAQ) – Saegis Campus, Senior Lecturer / Programme Coordinator - Institute of Human Resource Advancement(IHRA) -the University of Colombo, Senior Lecturer/Head of Division- National Institute of Business Management, Senior Lecturer/Lecturer /Assistant Lecturer- University of Colombo School of Computing, Graduate Project Assistant University of Colombo School of Computing. Mr. Ginige is a Member of the Institute of Electrical and Electronics Engineers (MIEEE), a Member of the British Computer Society (MBCS), Member of the Sri Lankan Computer Society (MCSSL). Session Chair if International Conferences, 2021 International Conference On Computer Communication And Artificial Intelligence (Ccai 2021) Guangzhou, China - Session 3: Computer Vision and Image Application, ICIM 2021 the 7th International Conference on Information Management, 2021 |London, UK. Co-Sponsored By Patrons EAET 2021, 2nd European Advanced Educational Technology Conference- Session 4: Information Teaching and Management, ICBT Annual International Research Symposium (AIRS 2018)- Co Char for Information Technology Session. Conference Committee Member as a Reviewer in International Conference on - ICGDA(January 21-23, 2022) Paris, France, 2nd European Advanced Educational-Technology Conference (EAET 2021, Imperial College, London, UK), 5th International Conference on Information System and Data Mining (ICISDM2021, Silicon Valley, CA, USA), 7th International Conference on Computing and Data Engineering (ICCDE2021, Phuket, Thailand), International Conference on Computer Communication and Artificial Intelligence (CCAI 2021, Guangzhou, China), 4th International Conference on Computers in Management and Business (ICCMB2021, Singapore), ICBT Annual International Research Symposium (AIRS 2020, 2019, 2018, Colombo, Sri Lanka)

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# A MULTI-INPUT MULTI-OUTPUT TRANSFORMER-BASED HYBRID NEURAL NETWORK FOR MULTI-CLASS PRIVACY DISCLOSURE DETECTION

A K M Nuhil Mehdy and Hoda Mehrpouyan

Department of Computer Science, Boise State University, Idaho, USA

## *ABSTRACT*

*The concern regarding users' data privacy has risen to its highest level due to the massive increase in communication platforms, social networking sites, and greater users' participation in online public discourse. An increasing number of people exchange private information via emails, text messages, and social media without being aware of the risks and implications. Researchers in the field of Natural Language Processing (NLP) have concentrated on creating tools and strategies to identify, categorize, and sanitize private information in text data since a substantial amount of data is exchanged in textual form. However, most of the detection methods solely rely on the existence of pre-identified keywords in the text and disregard the inference of underlying meaning of the utterance in a specific context. Hence, in some situations these tools and algorithms fail to detect disclosure, or the produced results are miss classified. In this paper, we propose a multi-input, multi-output hybrid neural network which utilizes transfer-learning, linguistics, and metadata to learn the hidden patterns. Our goal is to better classify disclosure/non-disclosure content in terms of the context of situation. We trained and evaluated our model on a human-annotated ground truth dataset, containing a total of 5,400 tweets. The results show that the proposed model was able to identify privacy disclosure through tweets with an accuracy of 77.4% while classifying the information type of those tweets with an impressive accuracy of 99%, by jointly learning for two separate tasks.*

## *KEYWORDS*

*Feature Engineering, neural networks, Natural Language Processing, Privacy.*

## 1. INTRODUCTION

Over the years with the increase in accessibility of internet and growth of communication platforms and social networking sites, user's concern about their privacy has also increased [31, 36, 52]. In order to provide usable tools and algorithms for users to manage the disclosure of their private information, much research has been carried out [37]. Mostly focused on understanding how users are sharing their private information through emails, text messages, and social media platforms and providing them with a clear picture of privacy threats and consequences of information sharing activities [11, 34].

Research in these areas is especially important, since the aggregated amount of personal information that an individual shares could be exploited by the modern AI (artificial intelligence) techniques to gain meaningful insights on their private information which could lead to serious privacy violations [20]. Wang at. al argues that user-specific targeted attacks are becoming more

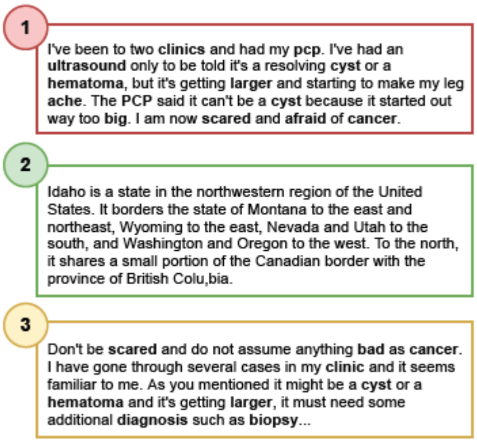
David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 221-241, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111419

common by exploiting the victim's private information [49]. Hence, the need to design and develop efficient tools and techniques to protect individual's privacy have resulted in researchers focusing on understanding the individual's motive to disclose private information [23, 32, 53].

Researchers in the field of Natural Language Processing (NLP) have concentrated on creating tools and strategies to identify, categorize, and sanitize private information in text data since a substantial amount of data is exchanged in textual form [2, 7, 42]. A usable privacy-disclosure detection tool is dependent on the understanding of what constitutes as private information and what defines a disclosure for an individual user. Different information is considered as private or sensitive across different domains of human lifestyle [8]. Researchers have also intended to classify someone's private information into two main categories: objective (i.e., factual information such as age, sex, marital status, health condition, financial situation) and subjective (i.e., internal states of an individual such as interests, opinions, feelings) [45]. As per the scope of this paper, we define *privacy disclosure* as an occurrence when a piece of text, which is usually a statement/expression from an author, contains someone's private information/situation. In other words, we focus mostly on the objective disclosure where users explicitly reveal someone's privacy. We consider three types of information disclosure in this research work: health condition, financial situation, or relationship issues.

For example, a disclosure occurs when a user tweets about his/her economic situation, i.e. the financial crisis he/she is going through, investment details, etc. Another example of disclosure could be when a patient tweets about his/her own physical/mental health condition, diagnosis results, medication/drug he/she is taking, etc. The intuition is similar for the Tweets that are about relationship issues. Likewise, we define non-disclosure as an event when a piece of text is not disclosing someone's health condition, financial situation, or relationship issues. Examples of non-disclosure information sharing activities are: when an activist tweets about the national/global financial crisis, observations about the stock market, tips and tricks for the new investors, etc. Another example of non-disclosure could be when a doctor tweets about a disease, its symptoms, health care advice, etc. Therefore, a usable privacy disclosure tool is required to differentiate between public/private information and overcome the difficulties associated with the natural language processing of context-based textual data.

As part of this efforts, a wide range of proposed methodologies such as dictionary utilization, information theory, statistical model, machine learning, and deep learning have shown promising results in identifying privacy disclosure in text data [7, 10, 18]. However, most of the methods are based on the fact that they solely rely on the existence of keywords/terms/phrases and disregard meaning inference from the text. We observed through our experimentation that these limitations, in some cases, result in miss classification. This is because only the existence of sensitive keywords in a piece of text does not always result in user's privacy disclosure.



1

I've been to two **clinics** and had my **pcp**. I've had an **ultrasound** only to be told it's a resolving **cyst** or a **hematoma**, but it's getting **larger** and starting to make my leg **ache**. The PCP said it can't be a **cyst** because it started out way too **big**. I am now **scared** and **afraid** of **cancer**.

2

Idaho is a state in the northwestern region of the United States. It borders the state of Montana to the east and northeast, Wyoming to the east, Nevada and Utah to the south, and Washington and Oregon to the west. To the north, it shares a small portion of the Canadian border with the province of British Colu,bia.

3

Don't be **scared** and do not assume anything **bad** as **cancer**. I have gone through several cases in my **clinic** and it seems familiar to me. As you mentioned it might be a **cyst** or a **hematoma** and it's getting **larger**, it must need some additional **diagnosis** such as **biopsy**...

Figure 1. Example of disclosure post (1), non-disclosure post (2), and highly similar to disclosure but actually a non- disclosure (3) [34].

For example, in Figure 1, the text from the box 1 is revealing someone's health information (i.e., the patient might have cancer) and the text from the box 2 is just an article about a state of the United States representing a public ambience. It is relatively an easy task in NLP to distinguish these two piece of texts based on the traditional techniques such as keyword spotting, bag-of-words model, rule based approach, etc. [5]. Now, the text from the box 3 contains similar keywords and sentence structure as the patient's post (box 1). However, this piece of text is not actually revealing the health information of the author. In other words, the doctor does not have cancer rather just a comment about cancerous disease. Therefore, it is quite challenging to distinguish the content of box 1 from the box 3, without taking into the consideration the underlying meaning and hidden patterns.

## 1.1. Contributions of the Work

We propose a novel and hybrid multi-input multi-output neural network based model that overcomes the NLP challenges by precisely identifying privacy disclosures through tweets by combining knowledge from pre-trained language model, semantic analysis, linguistics, and the use of metadata. The multi-input, multi-output model is able to identify both the information type (health, finance, relationship) of the tweets and the disclosure occurrence by jointly learning for two separate tasks. We also trained and evaluated our model on a human annotated ground truth dataset that contains a total of 5,400 tweets from anonymous users. Thus, our model could be implemented in the practical and usable fields of data privacy, information security, natural language processing, etc. A few of the notable contributions of this paper includes:

- Presenting a multi-input, multi-output hybrid neural network that utilizes pre-trained language model, and still make use of traditional linguistics and structured metadata.
- Evaluating its multi-output capability that jointly learns for solving two separate NLP tasks while utilizing a pre-trained language model.
- Sharing the model performance on a ground truth dataset for benchmarking.

The rest of the paper is organized as follows: section 2 contains the background and reviews on the related research works and a few of their limitations we observed. Section 3 describes about the dataset used in this paper along with the detail of the data labelling strategy. The methodology, data pre-processing, and feature engineering techniques are described in detail in section 4. The detail of the deep neural network architecture is presented in section 5 following the experiments in 6. Lastly, section 7 represents the experimental results following the conclusion.

## 2. BACKGROUND AND RELATED WORK

In this section, we review the state-of-the-art natural language processing research that are focused on privacy disclosure detection [3,8,22,33].Traditional research in this field has mostly relied on lexicon-based techniques to automate the content analysis of privacy-related information by leveraging linguistic resources such as privacy dictionaries. Existing automated content-analysis tool such as LIWC[1] is used with a specific sets of privacy dictionaries. Vasalou et al. suggest such a method that utilizes a dictionary of individual keywords or phrases which are previously assigned to one or more privacy domains [47]. To create the dictionary, they sample from a wide range of privacy domains such as self-reported privacy violations, health records, social network sites, children's use of the Internet, etc. However, their technique solely relies on the predetermined sensitive keywords/terms which classify both a medical article (public) and someone's medical condition (private) as a private document. Similar approach was taken by Chakaravarthy et. al. for a document sanitization task, where they represent a scheme that detects sensitive information using a database of entities [7]. The database contains different entities i.e. persons, organizations, products, diseases, etc. Each entity is also associated with a set of sensitive terms e.g. name, address, age, birth date, etc. Thus, a set of terms is considered as the context of the entity. For example, the context of a person become his/her age, birth date, name, etc.

Researchers from the area of information theory leverage large corpus of words along with computational linguistics to identify sensitive information in text documents [42]. Information theory provides the necessary formula for calculating the sensitivity score, otherwise known as IC (Information Content) score of every term, based on the amount of information it contributes to a corpus. For example, in a database of employee, a term such as *handicapped* carries more information than the common terms such as job, manager, desk, office, etc. All such terms that exceed a threshold score βare considered as sensitive. One of the advantages of this technique is that a finite collection of named entities is not required for the disclosure detection to be successful. However, this approach suffers from the same limitation as the previous line of work. In other words, it does not consider any semantic information other than merely relying on the appearance of sensitive keywords. Other popular techniques such as Named Entity Recognition (NER), also known as entity chunking, entity identification, or entity extraction have also been used by many researchers to identify and classify private information in text documents [2]. This line of research is based on the sub-task of information extraction technique that aims to identify named entities (medical codes, time expressions, quantities, monetary values, etc) and classify them into predefined categories in an unstructured text. Modern NER systems use linguistic grammar-based techniques, statistical models, machine learning, etc. Regardless of the underlying method, the NER based disclosure detection techniques also lack the capability of properly inferring the meaning from a text that could disclose someone's private information if a specific named entity is not detected (see examples 3 in Table 1).

Machine learning based techniques such as association rule mining [10], support vector machines (SVM), random forests [45], boosted Naive Bayes, AdaBoost, latent Dirichlet allocation (LDA), etc. have also been used to tackle similar tasks. Hart et al. used a novel training strategy on top of SVM to classify text documents as either sensitive or non-sensitive [18]. Caliskan et al. proposed a method for detecting whether or not a given text contains private information by combining

[1] Linguistic Inquiry and Word Count

topic modelling, named entity recognition, privacy ontology, sentiment analysis, and text normalization technique [6]. A combination of linguistic operations and machine learning is proposed by Razavi et. al. to detect health information disclosure [38]. They first compile a list of keywords related to a person's health information, and then apply keyword combinatorial web search. Alongside, they implement a machine learning layer to detect and learn any possible latent semantic patterns in the annotated dataset. Mao et al. studied privacy leaks on Twitter by automatically detecting vacation plans, tweeting under influence of alcohol, and revealing medical conditions [33] As the classifier model, they implemented two machine learning algorithms; Naive Bayes and SVM based on the TF-IDF (Term Frequency Inverse Document Frequency) feature space. Their main research goal was to analyse and characterize the tweets in terms of who leaks the information and how. Therefore, in the paper, the focus was less on the architecture and performance of the disclosure detection model.

Bak et. al. applied a modified LDA based topic modelling technique for semi-supervised classification of Twitter conversations that disclose private information [4]. This technique is also based on the distributions of terms/keywords across documents and corpus, which again does not consider word meaning inference. Most of the above-mentioned approaches have drawbacks since they rely exclusively on the presence of keywords and ignore word meaning inference from the text. We observe through our experimentation that, these limitations, in some cases, result in miss classification. This is because, existence/lack of sensitive terms/keywords in a piece of text does not always result in disclosure/non-disclosure of private information (see examples 3,4,5 in Table 1).

In order to overcome these limitations, recent research works from the area of NLP and privacy have considered utilizing semantic meaning along with lexical and syntactic analysis, while designing and developing deep learning based models [12,34,35,46]. Accordingly, there have been a significant progress in the area of language modelling through training complex models on enormous amounts of unlabelled data [14,48]. All the tailored solutions are being outperformed by this generic models. Most importantly, the utilization of transfer learning and pre-trained model have shed light into this area of research. Dadu et. al. proposed a predictive ensemble model by exploiting the fine-tuned contextualized word embedding, RoBERTa (Robustly Optimized BERT Approach) and ALBERT (A Lite version of BERT). The authors generated a small, labelled dataset, containing Reddit comments from casual and confessional conversations. Through the ensemble implementation they achieved 3% increment in the F1-score from the baseline model. Therefore, after considering the importance of transfer-learning and also taking into account the significance of linguistic features, we propose a multi-input hybrid neural network which utilize both transfer-learning and linguistics along with the metadata from the input text. The multi-output model is also able to classify both the information type of the input text and the disclosure occurrence by jointly learning for two separate tasks. We next describe the proposed framework.

## 3. Dataset

The deep learning based methodology proposed in this paper consists of a supervised neural network model that requires labelled data to learn the patterns of the disclosure and non-disclosure texts. There might be several reasons why no dataset is available for this purpose in literature, i.e., the restricted access policies of such data sources (e.g., emails, SMS, chat records), lack of privacy preserving research strategies, the complexity associated with the data labelling technique, etc. Therefore, we collected, and human annotated a ground truth dataset that contains human expressions, comprised of multiple English sentences, through which their privacy might have been disclosed. The following two sections detail our data collection and data labeling steps.

## 3.1. Data Collection

In order to collect diverse and user-centric data from different domains, we use the online platform, Twitter. People tend to prefer this platform to share their personal opinions, perceptions, issues, and observations through tweets which are comprised of a few sentences, hashtags, and emojis. We utilized Twitter search API [44] for mining the required dataset following a set of cleaning and labelling processes. We limited the data collection to those tweets that are written in English language and from anywhere in the world. This allows us to collect a generalized set of data written in different styles. The dates for crawling the tweets are randomly chosen for better sampling. Most importantly we filtered out the tweets based on a set of criteria such as i) tweets that contain any links, ii) retweets iii) replies to the tweets iv) tweets that are from verified accounts v) tweets that are posted by bots.

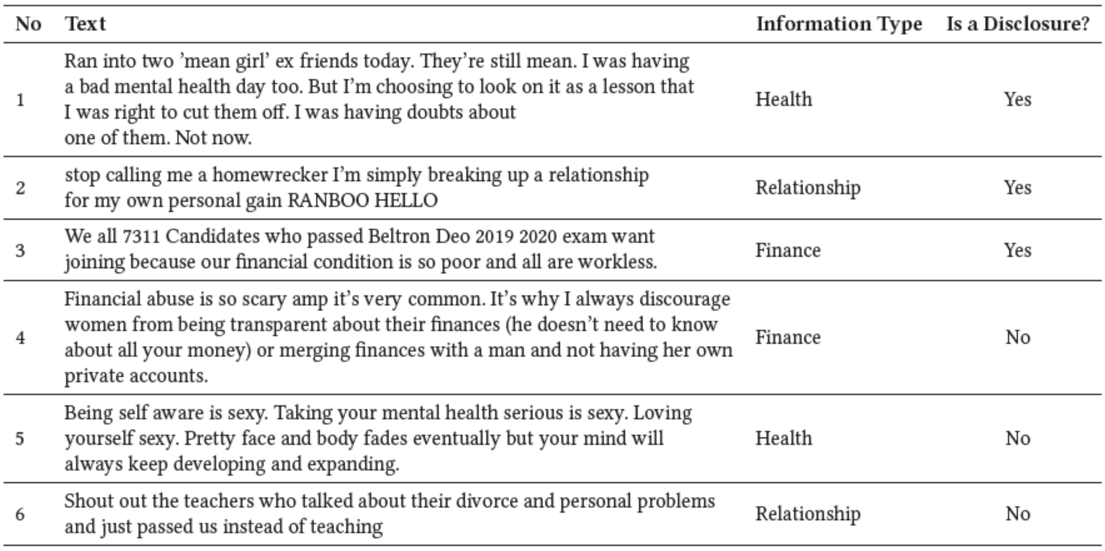
A total of 45,000 tweets is collected from three different privacy domains i) health, ii) finance, iii) relationship. The advanced search query strategies offered by the Twitter API [44] allowed us to properly identify and collect the tweets from these three categories. From these sets of tweets, we sampled a set of 6,000 random tweets based on the stratification on these three information types, selecting 2000 tweets from each category. This smaller subset of dataset is then used for human annotation and model training. In addition, we maintained the anonymity of the tweets by removing all the metadata excepts the tweet's date-time, tweet texts, and device-type used to post these tweets. Therefore, usernames, handles, permalinks, or tweet id remained hidden from the human annotators as an ethical consideration. We also meet the Twitter Developer Agreement and Policy[2].

## 3.2. Data Labelling

In each of the collected tweets, people tend to share their personal issues, opinions, perceptions, and advice, etc. It is observed that the authors intentionally or unintentionally disclose their own or someone else's private information such as health condition, financial situation, or relationship issues through their tweets. Some examples of such privacy disclosure and non-disclosure tweets can be found in Table 1 which are randomly sampled from the 6K dataset.

[2] "You may use the Twitter API and Twitter Content to measure and analyze topics like spam, abuse, or other platform health-related topics for non-commercial research purposes by conducting only non-commercial research on this dataset."

Table 1. Example of disclosure and non-disclosure tweets
(Samples are taken from the set of 6,000 tweets).



| No | Text | Information Type | Is a Disclosure? |
|---|---|---|---|
| 1 | Ran into two 'mean girl' ex friends today. They're still mean. I was having a bad mental health day too. But I'm choosing to look on it as a lesson that I was right to cut them off. I was having doubts about one of them. Not now. | Health | Yes |
| 2 | stop calling me a homewrecker I'm simply breaking up a relationship for my own personal gain RANBOO HELLO | Relationship | Yes |
| 3 | We all 7311 Candidates who passed Beltron Deo 2019 2020 exam want joining because our financial condition is so poor and all are workless. | Finance | Yes |
| 4 | Financial abuse is so scary amp it's very common. It's why I always discourage women from being transparent about their finances (he doesn't need to know about all your money) or merging finances with a man and not having her own private accounts. | Finance | No |
| 5 | Being self aware is sexy. Taking your mental health serious is sexy. Loving yourself sexy. Pretty face and body fades eventually but your mind will always keep developing and expanding. | Health | No |
| 6 | Shout out the teachers who talked about their divorce and personal problems and just passed us instead of teaching | Relationship | No |

We recruited human annotators from Amazon Mechanical Turk[3], an online crowd-sourcing marketplace to label all of the tweets as either disclosure or non-disclosure. The detailed instructions along with a set of good and bad examples of labelling was provided to assist the annotators understand the task correctly. We specifically guided them to follow the definitions of disclosure and non-disclosure, provided in section 2. We limited the selected annotators to USA with a good reputation (i.e., at least 95% HIT[4] approval rate) and those who are at least 18 years old. Each annotator was paid $0.05 per tweet based on our pilot trials indicating workers could label each tweet within 30 seconds. It is worth mentioning that only the binary labelling of disclosure/non-disclosure was completed by the human annotators. They were not asked to label the information types, since we already assigned these labels as a bi-product while crawling the tweets using the advanced search query API of Twitter. Most importantly, we employed 3 human annotators per tweet to decide whether or not that post is a privacy disclosure. This enabled us to select the most voted label for the tweet as the ground truth.

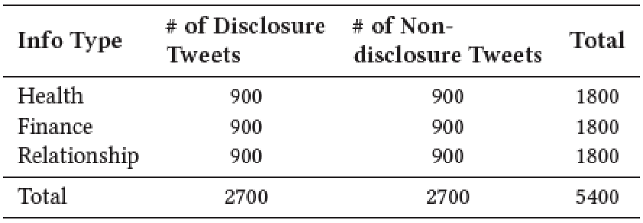
## 3.3. Data Augmentation

We discovered a moderate level of data imbalance after annotating the dataset. A total of 807 tweets out of 2000 from the health category and 769 tweets out of 2000 from the finance category were labelled as disclosure class whereas 799 tweets out of 2000 from the relationship category were labelled as non-disclosure. Therefore, we performed a data augmentation step to make the dataset balanced. First, we randomly sampled the candidate tweets to be augmented from each category. We sampled 93 disclosure tweets from the health category, 131 disclosure tweets from the finance category, and 101 non-disclosure tweets from the relationship category. Then we applied *domain-specific paraphrasing and synonym replacement* technique on these tweets as our augmentation strategy. This simple yet effective approach of augmenting text data has been recommended by the researchers and proved to be useful for getting generalized text data [50]. After the augmentation, we got 900 (800+93) disclosure tweets for the health category, 900

[3] A crowd sourcing website for businesses and researchers to hire remotely located "crowdworkers" to perform on-demand tasks such as survey, data labelling, etc.
[4] Human Intelligence Task

(769+131) disclosure tweets from the finance category, and 900 (799+101) non-disclosure tweets for the relationship category. On the other hand, we re-sampled 900 non-disclosure tweets from the health category, 900 non-disclosure tweets from the finance category, and 900 disclosure tweets from the relationship category. This resulted in a balanced dataset of 5,400 tweets where each of health, finance, and relationship categories contained 900 disclosure and 900 non-disclosure tweets (Table 2).

Table 2. Final dataset (balanced) for model training.



| Info Type | # of Disclosure Tweets | # of Non-disclosure Tweets | Total |
|---|---|---|---|
| Health | 900 | 900 | 1800 |
| Finance | 900 | 900 | 1800 |
| Relationship | 900 | 900 | 1800 |
| Total | 2700 | 2700 | 5400 |

## 4. METHODOLOGY

The neural network based model proposed in this paper adopts a transformer based pre-trained model called BERT (Bidirectional Encoder Representations from Transformers). We use this state-of-the-art pre-trained model to develop our custom multi-input multi-output model because: i) it supports fine-tuning for custom NLP tasks (transfer learning), ii) it is trained on a huge corpus of unlabelled texts (3,300 millions of words), ii) contains millions of parameters (110M), iv) supports parallelization for hardware acceleration, etc. The following subsections further detail on this component along with the data pre-processing and feature-engineering steps.

### 4.1. Data Pre-processing

As depicted in Table 1 both disclosure and non-disclosure tweets could contain similar keywords, sentence structure, and other syntactic constructs. This makes the classification problem particularly challenging, because we cannot simply rely on the lexical items and obvious keywords in the text, like bag-of-word models. Rather, we are required to discover the hidden patterns and infer author’s intentions that are embodied in the text, and to encode the underlying meaning expressed in the text to better classify the disclosure/non-disclosure activities. Therefore, unlike the traditional approaches that are mostly based on the bag-of-words technique, we kept the punctuation and stop words in the text to preserve the syntactic structure. We use NLP Toolkit [19] to clean the tweets in a customized way that ignores noisy and redundant tokens such as *‘‘,,’‘*, *‘‘;--’‘*, *‘‘!!!’‘*, *‘‘:-)’‘* and preserves the non-redundant ones such as *‘‘,’‘*, *‘‘;’‘*, *‘‘:’‘*, *‘‘.’‘*, *‘‘he’‘*, *‘‘the’‘*, *‘‘in’‘* etc. This is in contrast to the traditional approach of text analysis that is based on removing all the punctuation. It is important to note that we also removed Twitter specific tokens such as @, # and non-unicode special characters which might have been added by the users’ device.

### 4.2. Feature Engineering

We performed feature engineering on the dataset to produce four new features which then were fed into the neural network through its multiple input channels. Based on the Dependency Parse (DP) tree information of the texts, the underlying syntactic relationship of the data was generated. Additional features, i.e. date and time of the tweets, and the type of device that was used to post the tweets are also fed into the network as a meta data. Below we explain these new synthetic features in more details.

#### 4.2.1. Syntactic Structure

Certain formal properties of the language such as dependency parse tree information are known as a ‘‘purely stylistic’‘ by the theoretical linguistics [4]. In other words, two English sentences might have different syntactic forms but still express the similar meaning or vice versa [15]. For example, (*I suffered a lot in last few days*) with the DP structure *nsubj ROOT det dobj prep amod amod pobj* could be semantically equivalent to another sentence (*In last few days I suffered a lot*) having the structure *prep amod amod pobj nsubj ROOT det npadvmod*, though they are syntactically different. Figure 2 depicts the DP information of an example sentence where the DP tags are shown on the edges.

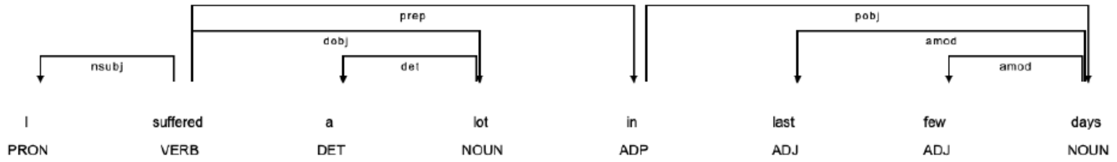


Figure 1. Dependency Parse Tree Information of a Sentence.

Along with the parts of the speech tags, these types of representation of the language features enable the deep-learning based models to learn about the sequential patterns of the sentence constructs along with the arrangements of the word token themselves [35]. This helps the model. Hence, we used a natural language toolkit [19] to extract the information to enrich the feature space of the dataset.

### 4.3. Transfer Learning and Fine Tuning

Most of the NLP tasks such as text classification, machine translation, text generation, language modelling, etc. are considered as sequence modelling tasks. Typical machine learning models such as bag-of-words, term-frequency inverse document-frequency, and multi-layer perception are not able to capture the sequential information presented in the text. Therefore, to capture this important piece of information, researchers have introduced techniques such as recurrent neural network (RNN) and long short-term memory-based network. However, these types of neural networks introduce new issues in terms of performance and efficiency. For the reason that both RNN and LSTM based neural network takes one input (token in case of text sequence) at a time, they could not be parallelized. This makes the training operation, time consuming, specially while handling a large dataset.

This was the case until 2018 when Google introduced the transformer model which turned out to be ground-breaking [48]. It is mainly an attention mechanism for learning contextual relations between words in text (Figure 3).

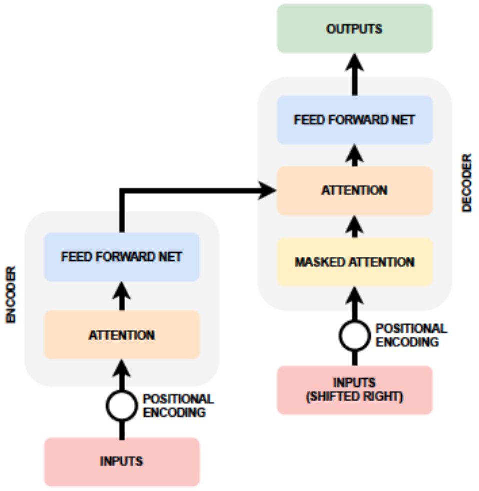


Figure 2. Simplified View of the Transform Architecture [48]

It also introduced an architecture that supports parallelization and make use of unlabelled text data for training. In the following year, BERT has been introduced which makes use of the transformer architecture. It is a new language representation model published by the researchers from the Google AI Language team in 2018 [14]. Since then all the tailored solutions to various NLP tasks are being outperformed by this generic transformer based model. Most importantly, BERT supports transfer-learning which allows us to develop domain specific custom NLP models while utilizing the power of transformer based pre-trained models. Transfer learning is pre-training a neural network model on an informed task and then using the trained network as the basis of a new purpose-specific model, otherwise known as fine-tuning [43]. Researchers from the area of computer vision have already shown the significance of this technique [17], and in recent years, they have been showing how a similar technique could be useful in natural language tasks as well [40]. Figure 4 depicts an abstract view of BERT's pre-training and fine-tuning Procedures.

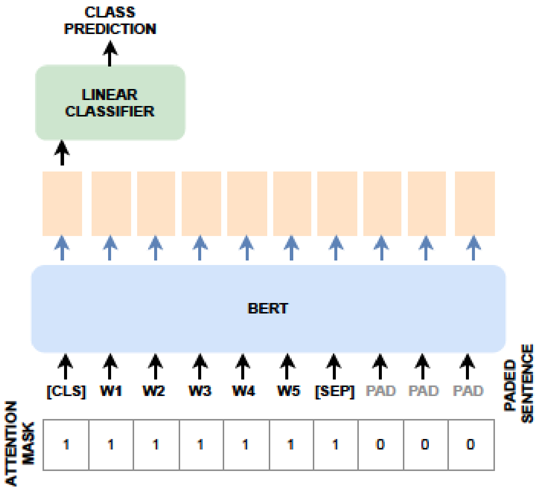


Figure 3. Simplified View of BERT Fine-tuning Procedures [14]

Therefore, we adopt transfer-learning technique to design and develop our hybrid multi-input multi-output neural network which not only fine-tune a pre-trained model but also make use of the linguistic pattern-learning and metadata utilization. There are different ways of fine tuning a model: i) the entire architecture could be further trained on a new dataset which allows the model to update its pre-trained weights ii) retraining only the higher layers while keeping the weights of initial layers of the model frozen iii) keeping the all the layers of the model frozen, and add one or more new neural network layers of our own, where only the weights of the new layers will be updated during the training phase. In this paper, we utilize the last technique where we import the pre-trained BERT model as a neural network layer into our custom neural network architecture. This acts as one of the three main input channels of our network. The other two input channels provide additional data to the model that we detail in the following sections.

% Using transformer-based models has multiple advances including but not limited to - i) they can take the entire sequence of tokens as text input enabling the capability of training acceleration by GPUs and TPUs, ii) no need of labelled data for pre-training the model, iii) they are better for transfer learning iv) supports better model explainability. Most significantly, this pre-trained model can be fine-tuned with just one extra output layer to produce state-of-the-art models for a wide range of applications with little task-specific architectural changes. Having the ability to be fine-tuned is also advantageous because these types of models, for example BERT, consists of a huge number of parameters (100M - 300M). Therefore, training such a model from the scratch on a relatively smaller dataset may result in poor performance (e.g., over-fitting or under-fitting).

## 5. Neural Network Architecture

The architecture of the proposed neural network is divided into three main segments: i) pre-trained BERT model ii) implementation of the linguistic features iii) integration of structured metadata. The output of this model consists of two different branches: i) multi-class classification of information types ii) binary classification of disclosure/non-disclosure information sharing transactions. Figure 5 depicts the architecture of the proposed multi-input, multi-output hybrid neural network. In the following subsections, we describe each component of the model in detail.

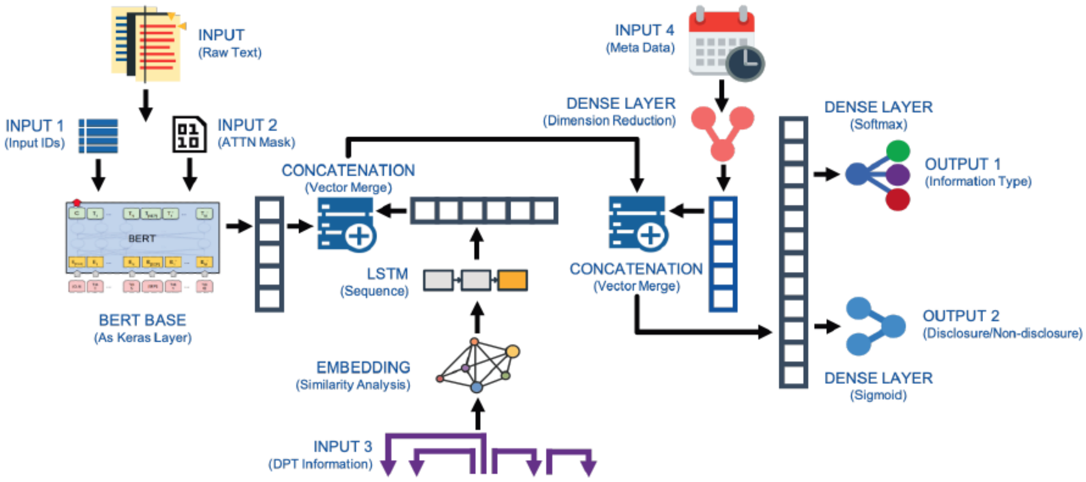


Figure 4. Architecture of the Proposed Model

### 5.1. Leveraging BERT

BERT model has two inputs: first from the word tokens, and second from the segment layer following their embedding layers. BERT has a vocabulary of 30,000 distinct tokens comprised of complete English words and word piece components (e.g., embedding for both *play* and *##ing* to work with *playing*). These tokens are associated with an initial embedding space known as WordPiece embedding. The two inputs are added and summed over a third embedding known as position embedding, followed by the dropout layers and layer normalization. The resulting BERT model contains 12 multi-headed self-attention layers (encoders) which are identical to each other. BERT is trained on two NLP tasks: i) the Next Sentence Prediction (NSP), ii) Masked Language Modelling (MLM). These two tasks are informally called fake tasks. In other words, when the pre-training of BERT happens, the model learns the language patterns while solving for these two given tasks. In the end, the trained model is saved and used for further fine-tuning to solve specific NLP tasks, like one in this paper (disclosure and information type classification). Please refer to the original research paper for a detailed implementation of the BERT's architecture [14].

### 5.2. Inputs to the Proposed Network

As mentioned previously, we import the pre-trained BERT model as a layer into the proposed neural network. Token-ids and attention-masks are fed into the model through two input channels of the BERT layer. Token ids are the integer encoded values for each of the tokens of the input text. Attention masks are supporting vectors that enable BERT differentiate between the actual and padding tokens. We add a dropout layer after the BERT main layer as suggested by the literature [21]. In addition, a separate input channel was added to the proposed neural network through which we fed the dependency parse tree information of the same input text. This input path has its own embedding layer which gets learned during the training process. Then we added an LSTM layer that learns the sequential information of the dependency tags. Output of this LSTM layers are then concatenated with the output of the dropout layer. At this stage, we employ the metadata to the neural network through another input channel. This input takes the day of the week, hour of the day, and device type information associated with each input text. A dense layer is added to reduce the dimensionality caused by the encoding of these categorical features. This dense layer uses rectified linear unit as its activation function. Finally, we concatenated the output of this input channel with the output of the previous concatenation operation (BERT's output + DP output).

### 5.3. Outputs from the Proposed Network

Since we aim to solve two parallel tasks through a single neural network model, there are two separate output layers in the proposed model. In one output layer we add three neurons that result a probability distribution of the information type variable. The predicted probabilities of an input text being any of the three classes: health, finance, relationship is distributed among these three neurons. The neuron with the highest probability wins and shows the information type of the input text. The other output layer is comprised of a single neuron which calculate the probability of the input text being either disclosure or non-disclosure. In other words, the model jointly optimizes for a multi-class classification task and a binary-class classification task. Therefore, we employ different loss functions for these two separate out layers. The multi-class prediction layer uses categorical cross entropy, and the binary class prediction layer uses binary cross entropy with accuracy as the evaluation metrics.

## 6. EXPERIMENTS

In this section we describe the implementation detail of the proposed neural network architecture along with the tools we used. We also talk about the optimizer, loss functions, metrics, and a set of hyper-parameters in this section.

### 6.1. Tools and Libraries

We utilize the Huggingface's Transformers package which is an open source natural language processing library developed in Python programming language [51]. This library lets developers import a wide range (32+ pre-trained models in 100+ languages) of transformer-based pre-trained models such as BERT, ALBERT, XLnet, GPT-2, etc. It is also very easy to switch between different transformer based models through Huggingface Transformers. Most importantly, it supports interoperability between PyTorch, TensorFlow, and other deep learning libraries. We use Tensorflow that comes with Keras pre-built to architect the multi-input multi-output neural network [1]. More specifically, we use the the Keras functional API to create the neural network architecture [26].

We make use of the *TFBertModel* module from the Transformers package which is an interface to the Tensorflow library. We import the pre-trained BERT model called *bert-base-uncased* using this module. This is a pre-trained model on English language, and it is uncased meaning it does not make a difference between the words playing and Playing [13]. This specific base model consists of 110 million parameters. The main layer of this pre-trained model is imported as a keras layer into our custom architecture following a dropout layer. In the other input channel of our model, a LSTM layer with *tanh* activation function is used over the dependency parse tree information by utilizing the keras *LSTM* layer [27]. Before this layer, we use the keras *Embedding* layer to learn the embedding of these dependency tags in a 16-dimensional vector space [25]. The Keras *concatenate* method then takes the output from this LSTM layer and the dropout layer from BERT to merge them into a single vector. The final input into our custom neural network makes use of a keras *Dense* layer [24], and its output is also gets concatenated with the other branch before going through the final output layers.

For text pre-processing, we applied Spacy [19] to derive the dependency parse tree information of each tweet. Spacy provides dependency parser, trainable models, tokenizer, noun chunk separator, etc. in a single toolkit. It offers the fastest syntactic parser in the world and its accuracy is within 1% of the best available natural language toolkit [9]. To perform the data augmentation step, we used another Spacy based library called spaCy WordNet [39]. It is a custom component for using WordNet and WordNet domains with spaCy which allows users to get synsets for a processed token filtering by domain. Text encoding and padding for these tag based sequences are done using Keras text to sequence and padding methods respectively [28]. To tokenize, pad, and prepare the raw texts for the BERT side input, we utilize the *BertTokenizerFast* that comes with the Transformer package. This tokenizer converts the raw texts into BERT compatible format such as adding special tokens ([CLS], [SEP]), truncating longer sequences, returning token ids and attention masks, etc.

### 6.2. Optimizer, Loss, and Metrics

We use *Adam* gradient descent algorithm as the optimization method for the neural network. It is considered to be computationally efficient and has little memory requirement [30]. The separate output heads use two different logarithmic loss functions: categorical cross entropy for information type classification, and binary cross entropy for the disclosure detection. The network uses *accuracy* as the optimization metrics for both of the output heads which is evaluated by the model during training and testing.

### 6.3. Hyper-Parameters

In case of fine-tuning based training, most of the hyper-parameters of the core model itself stay the same. Therefore, we also retain the hyper-parameters of BERT as it is. However, readers can refer to the BERT paper which gives specific suggestions on the hyper-parameters that require further tuning. In this section we only describe the about those hyper-parameters which we use for our custom neural network model.

First of all, we consider 55 (mode) as the maximum length of the input text sequences. Since the tweets in the dataset are of varying length, we use truncation and padding to make all the tweets have this same length. The first custom input that takes the dependency parse tree information learns an embedding space of length 16 with a vocabulary size of 47. The subsequent LSTM layers is comprised of 32 units which is the dimensionality of its output space. This layer uses *tanh* as the activation function with no *dropout*. All other parameters are kept default from the keras implementation [27]. The Keras *concatenate* method takes the output from this LSTM layer (32 dimensions) and the dropout layer from BERT (768 dimensions) to merge them into a single vector of 800 dimensions. The other custom input channel (metadata input) uses a dense layer with 32 neurons and rectified linear unit as their activation function which reduces its 149 dimensional input to 32. One of the final output layers that classify the information type uses 3 neurons with *softmax* activation. The other output that detects disclosure uses a single neuron with *sigmoid* activation. Both of these output layers use truncated normal distribution as the kernel initializers where the standard deviation is 0.02 for initializing all weight matrices. This value comes as default from the standard implementation of BERT by the Transformer library. The parameters for the Adam optimizer are chosen as follows: learning rate = *5e-04*, epsilon (a small constant for numerical stability) = *1e-08*, clipnorm (gradient norm scaling) = *1.0*. Other parameters of this optimizer are kept as default from the Tensorflow implementation it [29].

The whole dataset is split into a 90-10 ratio for training and testing respectively. For the validation purpose, we kept 20% from the training dataset while the model training process happens. Thus, 10% of the original dataset are used as test dataset which was never shown to the model. We feed the input data to the model with a batch size of 64, and let the model train for 5 epochs. We achieved the best performance from the model withing this amount of iterations. It’s worth mentioning that, all the above mentioned hyper-parameters are chosen based on several trials and outcomes.

### 6.4. Computing Resources

We used Google Colaboratory [16] as the experimentation platform which provided a Nvidia Tesla T4 GPU with 16GB memory. It took 15 minutes in average to run a complete training phase given the hyper-parameters that we mentioned already. Since this platform provides virtual infrastructure and sometime shares the resources among the users, the reported time may vary.

## 7. RESULTS

The results show that, by utilizing transfer learning and pre-trained language model, a multi-input neural network based model can be trained that learns beyond simple keyword spotting and utilizes linguistic features to classify whether or not a piece of text contains a privacy disclosure with a useful degree of accuracy. Moreover, through the experimentation, it is observed that, integration of metadata to the model increases the performance noticeably (increasing the accuracy by 1.80%). Since our dataset is balanced, we report receiver operating characteristic (ROC) curve, precision[5]. and recall[6]. score, f1-score[7], confusion matrix, and accuracy[8] score for both the binary and multi-class classification task.

## 7.1. Evaluation Considerations

The classification of *information type* is not the main and only evaluating pillar of the proposed model; rather, the classification of the *disclosure vs. non-disclosure* text is the main focus of the paper. In other words, our multi-input multi-output model is designed to solve the challenging task of distinguishing highly similar texts into disclosure and non-disclosure class. The *information type* classification is a bi-product while jointly training the multi-output model. Also, since we collected the tweets from three different information domains by utilizing the Twitter API, the texts are already well-aligned with these three classes. Therefore, this is expected to achieve a higher degree of accuracy while classifying the information types. However, classifying those tweets as disclosure vs. non-disclosure becomes a crucial challenge to solve. It can also be observed from Figure 6, which depicts how various models struggle to achieve better accuracy on the binary classification task. This is expected, as we described earlier, the textual similarities between these two classes of texts. It is also evident that a binary classification task on top of highly similar texts is still a challenge [12].

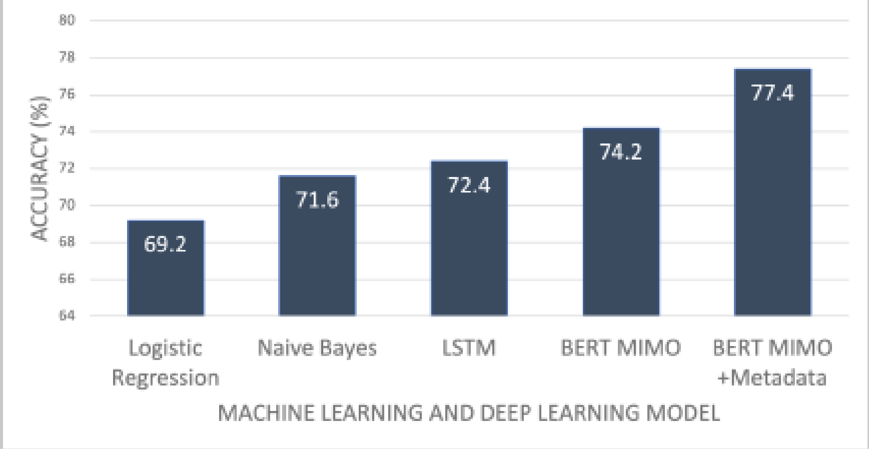


Figure 5. Comparison among the binary classification (dis- closure vs. non-disclosure) models.

In Table 3 and Table 4, we describe the classification report for both information type and disclosure detection respectively. As can be seen from these tables, the information type classifier achieves an impressive accuracy of 99%. The disclosure/non-disclosure classifier reaches up to 77.4% which is 8.2% more than bag-of-words and RNN based baseline models. We can also see a good recall score for the binary classifier which depicts its capability to detect most of the disclosure texts. In other words, 77% of all the disclosure texts have been identified successfully.

Table 3: Classification report for Disclosure/non-Disclosure on the Test Dataset (10% of 5400=540)

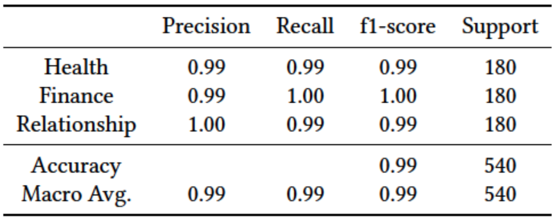
[5] What fraction of predictions as a positive class were actually positive.

[6] What fraction of all positive samples were correctly predicted as positive.

[7] The harmonic mean (average) of the precision and recall.

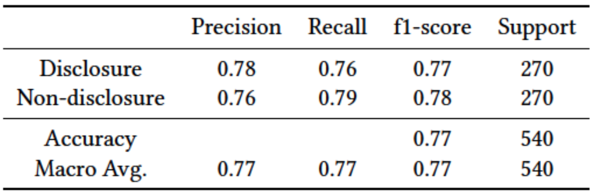
[8] The fraction of the total samples that were correctly classified.

Table 4: Classification report for Disclosure/non-Disclosure on the Test Dataset (10% of 5400=540)



| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Health | 0.99 | 0.99 | 0.99 | 180 |
| Finance | 0.99 | 1.00 | 1.00 | 180 |
| Relationship | 1.00 | 0.99 | 0.99 | 180 |
| Accuracy | | | 0.99 | 540 |
| Macro Avg. | 0.99 | 0.99 | 0.99 | 540 |

Table 5: Classification report for Information Types on the Test Dataset (10% of 5400=540)



| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Disclosure | 0.78 | 0.76 | 0.77 | 270 |
| Non-disclosure | 0.76 | 0.79 | 0.78 | 270 |
| Accuracy | | | 0.77 | 540 |
| Macro Avg. | 0.77 | 0.77 | 0.77 | 540 |

Figure 7 depicts the confusion matrix for information type classification. It can be seen that, only a few miss-classifications have occurred specially when the information type of the texts was *Relationship*. Likewise, Figure 8 depicts the confusion matrix for disclosure/non-disclosure classification. In Figure 9, we show the ROC curve for information type classification, and in Figure 9 we show the ROC curve for disclosure/non-disclosure classification. The binary classifier shows an area under curve (AUC) score of 0.834. Unlike the binary class ROC curve, we render the multi-class ROC curve by using one-vs-all technique to properly represent its performance.

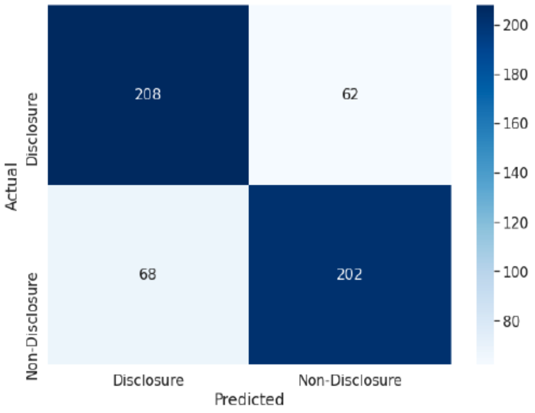


Figure 6. Confusion matrix for disclosure classification

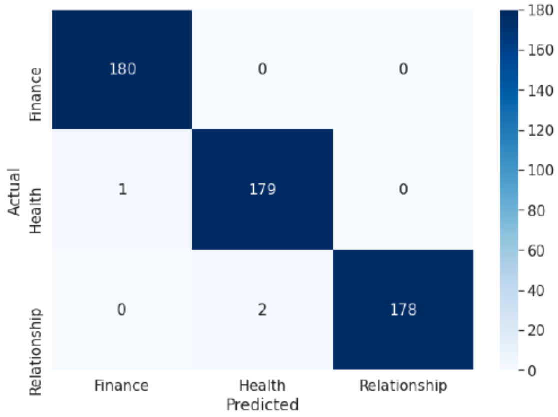


Figure 7. Confusion matrix for information type classification

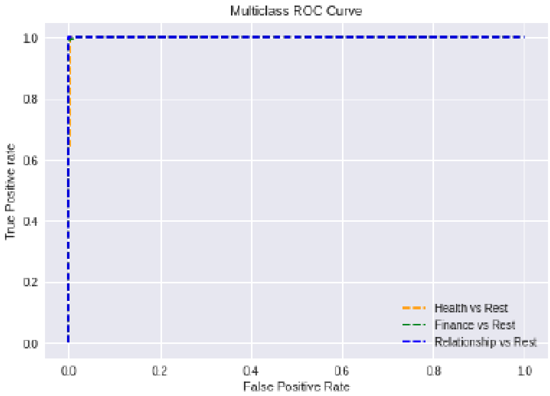


Figure 8. ROC Curve for Information Type Classification

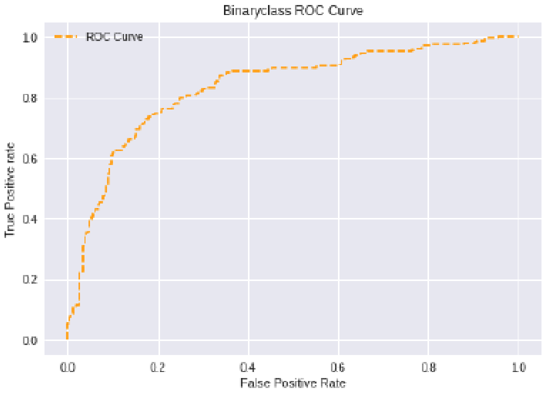


Figure 9. ROC Curve for Disclosure Classification

The performance of our model is not directly comparable with other similar approaches proposed in the literature because of the lack of common and shared dataset with similar properties. However, the closest and recent work of detecting self-disclosure on the #OffMyChest dataset, which contains Reddit comments, is worth comparing [12]. In their work, they achieved an accuracy of 74.12% and 74.20% on two different classes of the dataset: information disclosure and emotional disclosure respectively. Also, the precision and recall scores were 0.710, 0.551, and 0.636, 0.510 respectively. In comparison, the performance of our model is noticeably better in all the metrics.

## 8. CONCLUSION

In this paper we have proposed a multi-input, multi-output hybrid neural network that utilizes state-of-the-art transformer based pre-trained model called BERT along with language features and metadata to precisely detect privacy disclosure in text data. We also evaluate our model on a ground truth dataset that contains a total of 5,400 tweets from three different privacy domains: health, finance, and relationship. Unlike the traditional text classification techniques that primarily rely on keyword spotting, this model focus on underlying meaning and hidden patterns by leveraging pre-trained language model and classical linguistics. Additionally, our proposed architecture shows capability of solving two separate text classification tasks withing a single model that provides new insights which can help build practical NLP models. However, there are improvement scopes in the work presented in this paper. The learning and predictive performance of the model can be evaluated on a diverse dataset by taking samples from different data sources. Therefore, we want to collect a diverse dataset on various privacy domains in the future, using more sources such as forums, emails, text messages, and so on. In addition, performing privacy-preserving text analysis, and testing the integration of the model to the end products could also be future works. Most importantly, we plan to integrate explainability into the model for its fairness and trustworthiness.

## ACKNOWLEDGEMENTS

The authors would like to thank National Science Foundation for its support through the Computer and Information Science and Engineering (CISE) program and Research Initiation Initiative (CRII) grant number 1657774 of the Secure and Trustworthy Cyberspace (SaTC) program: A System for Privacy Management in Ubiquitous Environments.

## REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. 2011. On the declassification of confidential documents. In International Conference on Modeling Decisions for Artificial Intelligence. Springer, 235–246.

[3] JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 60–64.

[4] JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 1986– 1996.

[5] Roy F Baumeister and Kenneth J Cairns. 1992. Repression and self-presentation: When audiences interfere with self-deceptive strategies. Journal of Personality and Social Psychology 62, 5 (1992), 851.

[6] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In Proceedings of the 13th Workshop on Privacy in the Electronic Society. ACM, 35–46.

[7] Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. 2008. Efficient techniques for document sanitization. In Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 843–852.

[8] Dongjin Choi, Jeongin Kim, Xeufeng Piao, and Pankoo Kim. 2013. Text Analysis for Monitoring Personal Information Leakage on Twitter. J. UCS 19, 16 (2013), 2472–2485.

[9] Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 387–396.

[10] Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 893–901.

[11] Paul C Cozby. 1973. Self-disclosure: a literature review. Psychological bulletin 79, 2 (1973), 73.

[12] Tanvi Dadu, Kartikey Pant, and Radhika Mamidi. 2020. Bert-based ensembles for modeling disclosure and support in conversational social media text. arXiv preprint arXiv:2006.01222 (2020).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[15] David A Evans and Chengxiang Zhai. 1996. Noun-phrase analysis in unrestricted text for information retrieval. In Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 17–24.

[16] Google. [n.d.]. Colaboratory - Google Research. https://research.google.com/ colaboratory/ [Online; accessed 01-May-2021].

[17] Kasthurirangan Gopalakrishnan, Siddhartha K Khaitan, Alok Choudhary, and Ankit Agrawal. 2017. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Construction and Building Materials 157 (2017), 322–330.

[18] Michael Hart, Pratyusa Manadhata, and Rob Johnson. 2011. Text classification for data loss prevention. In International Symposium on Privacy Enhancing Technologies Symposium. Springer, 18–37.

[19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https: //doi.org/10.5281/zenodo.1212303

[20] Han Hu, NhatHai Phan, Soon A Chun, James Geller, Huy Vo, Xinyue Ye, Ruoming Jin, Kele Ding, Deric Kenne, and Dejing Dou. 2019. An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. Computational Social Networks 6, 1 (2019), 1–19.

[21] Huggingface. [n.d.]. Huggingface Transformers. https://huggingface.co/ transformers/main_classes/configuration.html Online; accessed 01-May-2021].

[22] Prateek Jindal, Carl A Gunter, and Dan Roth. 2014. Detecting privacy-sensitive events in medical text. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. 617–620.

[23] Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. Human–Computer Interaction 25, 1 (2010), 1–24.

[24] Keras. [n.d.]. Dense Layres - Keras Documentation. https://keras.io/api/layers/ core_layers/dense/ Online; accessed 01-May-2021.

[25] Keras. [n.d.]. Embedding Layres - Keras Documentation. https://keras.io/api/ layers/core_layers/embedding/ Online; accessed 01-May-2021.

[26] Keras. [n.d.]. Guide to the Functional API - Keras Documentation. https: //keras.io/guides/functional_api/ Online; accessed 01-May-2021].

[27] Keras. [n.d.]. LSTM layer - Keras Documentation. https://keras.io/api/layers/ recurrent_layers/lstm/ Online; accessed 01-May-2021.

[28] Keras. [n.d.]. Text Preprocessing - Keras Documentation. https://keras.io/ preprocessing/text/#tokenizer [Online; accessed 01-May-2021].

[29] Keras. [n.d.]. Text Preprocessing - Keras Documentation. https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam [Online; accessed 01-May-2021].

[30] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[31] Shulamit Sara Klinger. 2002. ‘‘ Are they talking yet?’‘: online discourse as political action in an education policy forum. Ph.D. Dissertation. University of British Columbia.

[32] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users’ information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information systems research 15, 4 (2004), 336–355.

[33] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. ACM, 1–12.

[34] AKM Nuhil Mehdy and Hoda Mehrpouyan. 2020. A User-Centric and Sentiment Aware Privacy-Disclosure Detection Framework based on Multi-input Neural Network.. In PrivateNLP@ WSDM. 21–26.

[35] Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. 2019. Privacy Disclosures Detection in Natural-Language Text Through Linguistically-Motivated Artificial Neural Networks. In International Conference on Security and Privacy in New Computing Environments. Springer, 152–177.

[36] Eni Mustafaraj and Panagiotis Takis Metaxas. 2011. What edited retweets reveal about online political discourse. In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.

[37] Melanie Nguyen, Yu Sun Bin, and Andrew Campbell. 2012. Comparing online and offline self-disclosure: A systematic review. Cyberpsychology, Behavior, and Social Networking 15, 2 (2012), 103–111.

[38] Amir H Razavi and Kambiz Ghazinour. 2013. Personal Health Information detection in unstructured web documents. In Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on. IEEE, 155–160.

[39] recognai. [n.d.]. spaCy WordNet. https://pypi.org/project/spacy-wordnet/ Online; accessed 01-May-2021].

[40] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. 15–18.

[41] Jacqueline Strunk Sachs. 1967. Recopition memory for syntactic and semantic aspects of connected discourse. Perception & Psychophysics 2, 9 (1967), 437–442.

[42] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2012. Detecting sensitive information from textual documents: an information-theoretic approach. In International Conference on Modeling Decisions for Artificial Intelligence. Springer, 173–184.

[43] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 242–264.

[44] Twitter. [n.d.]. Tweets Search API Reference. https://developer.twitter.com/ en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets Online; accessed 01-May-2021].

[45] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In The World Wide Web Conference. 3272–3278.

[46] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2020. A Study of SelfPrivacy Violations in Online Public Discourse. In 2020 IEEE International Conference on Big Data (Big Data). IEEE, 1041–1050.

[47] Asimina Vasalou, Alastair J Gill, Fadhila Mazanderani, Chrysanthi Papoutsi, and Adam Joinson. 2011. Privacy dictionary: A new resource for the automated content analysis of privacy. Journal of the Association for Information Science and Technology 62, 11 (2011), 2095–2105.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762 (2017).

[49] Ding Wang, Zijian Zhang, Ping Wang, Jeff Yan, and Xinyi Huang. 2016. Targeted online password guessing: An underestimated threat. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. ACM, 1242– 1254.

[50] Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 (2019).

[51] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019).

[52] Ekstrand, Michael D., Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In Conference on Fairness, Accountability and Transparency, pp. 35-47. PMLR, 2018.

[53] Mehrpouyan, Hoda, Ion Madrazo Azpiazu, and Maria Soledad Pera. Measuring personality for automatic elicitation of privacy preferences. In 2017 IEEE Symposium on Privacy-Aware Computing (PAC), pp. 84-95. IEEE, 2017.

## AUTHORS

**A K M Nuhil Mehdy** completed his Ph.D. degree from Boise State University, USA under the Cybersecurity emphasis, in Summer 2021. Earlier, he completed his B.Sc. Eng. degree in Computer Science back in 2011 from Rajshahi University of Engineering and Technology, Bangladesh and M.S. in Computer Science in 2017 from Lamar University, USA. His research interests include Privacy and Security, related to the internet users, Industrial Control Systems, Internet of Things, and Distributed Systems. Currently he is working as a Machine Learning Engineer at Micron Technology, Inc., USA.

**Hoda Mehrpouyan** is an associate professor of the Computer Science department at Boise State University, USA. Dr. Mehrpouyan's research focuses on ensuring privacy, security, and robustness of mission-critical cyber-physical systems. In May 2019, she was awarded a National Science Foundation CAREER Award from the Secure and Trustworthy Cyberspace (SaTC) program. Further, she received funding from the National Security Agency (NSA), Idaho National Lab (INL), and Idaho Secretory of State. She has more than 35 peer reviewed publication.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# Subtractive mountain clustering algorithm applied to a chatbot to assist elderly people in medication intake

Neuza Claro, Paulo A. Salgado, and T-P Azevedo Perdicoúlis

Escola de Ciências e Tecnologia
Universidade de Trás-os-Montes e Alto Douro, Vila Real 5000–811, Portugal

**Abstract.** Errors in medication intake among elderly people are very common. One of the main causes for this is their loss of ability to retain information. The high amount of medicine intake required by the advanced age is another limiting factor. Thence, the design of an interactive aid system, preferably using natural language, to help the older population with medication is in demand. A chatbot based on a subtractive cluster algorithm, included in unsupervised learned models, is the chosen solution since the processing of natural languages is a necessary step in view to construct a chatbot able to answer questions that older people may pose upon themselves concerning a particular drug. In this work, the subtractive mountain clustering algorithm has been adapted to the problem of natural languages processing. This algorithm version allows for the association of a set of words into clusters. After finding the centre of every cluster — the most relevant word, all the others are aggregated according to a defined metric adapted to the language processing realm. All the relevant stored information is processed, as well as the questions, by the algorithm. The correct processing of the text enables the chatbot to produce answers that relate to the posed queries. To validate the method, we use the package insert of a drug as the available information and formulate associated questions.

**Keywords:** chatbot, medicine intake aid system, natural language processing, subtractive mountain clustering.

## 1 Introduction

Physicians often prefer the treatment of illness using medication. One main reason for this is the non-invasiveness of this method of cure and another is the advance in science, namely in pharmacological engineering, that has given rise to new drugs and techniques with great effectiveness.

Older adults are more susceptible to the use of medication because they are also more prone to having chronic disorders such as high blood pressure, cardiac arrhythmia and diabetes. The use of multiple drugs to treat cumulative diseases — multiple pharmacy — and also the use of numerous medications to treat a single condition — polypharmacy — are very common situations among this age group. A study carried out in the period 2010–2011 in the United States [1] showed that almost 90% of elderly adults regularly take at least one prescription drug, approximately 80% take at least two, and 36% take at least five prescribed medications.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 243-258, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111420

With the upsurge in medication, errors associated with their use have also raised, especially in older people [2]. The error associated with taking drugs is especially problematic for these generations because, besides being the ones with the most complex clinical conditions, they also have cognitive problems related to memory and assimilation of information, making it difficult to take the right medication at the right time and in the manner prescribed by their doctor. It is estimated that close to half of the older adults do not take their medication according to the doctor's prescription [3]. Incorrect drug administration can happen by (i) taking medication at a different time of the day than it was prescribed; (ii) taking a different dose from the prescribed one; and (iii) changing the medication in course on a particular occasion [4].

These events may be caused by the similarity of the medicine box, the shape and colour of the pill, the similarity between the names of the medicines and the complexity and length of the medical prescription [4]. One study confirmed that 25% of the medication errors are associated with confusion with the name and 33% with confusion with the medication box and package insert [5]. Medication intake errors can lead to loss of treatment effectiveness and increased risk of new complications may induce a new disease state, a new hospitalisation or even death.

Technology information in the health area has been growing [6]. The use of systems to avoid medication errors at the hospital and home medication at home is vast. For example, Ahmed and coworkers [7] developed an automatic drug dispensing system. Mobile phone applications are also numerous [8] [9].

This work aims to develop an automatic conversational agent, a chatbot, capable of accompanying older people with taking medication. The person should be able to interact with this aid system in natural language.

The increase in the processing capacity of computers and smartphones, associated with the great development in artificial intelligence, has allowed the development of largescale software to facilitate the daily lives of humans. A bot, diminutive of a robot, is an automated hardware or software machine with the capacity to simulate human behaviour [10]. Bots are powered by advances in Artificial Intelligence (AI) technologies. Inside a bot, an algorithm can produce a certain answer according to the input data. An example of a bot is the chatbot, a program capable of having an online conversation with a human being.

The evolution in machine learning algorithms, such as deep learning and deep reinforcement learning, has improved natural language processing (NLP) performance. These advances have made intelligent conversational systems gain more and more popularity. Since then, chatbots have been applied in the most diverse areas: from commercial use [11] to medicine [12] [13].

Various NLP tasks are carried out to resolve the ambiguity in speech and language processing. Before machine learning techniques, all NLP tasks are carried out using various rules-based approaches. In rule-based systems, rules were constructed manually by linguistic experts or grammarians for particular tasks. Machine learning and statistical techniques are everywhere in today's NLP. In literature, the implementation of various machine learning techniques for various NLP tasks has been investigated extensively. The machine learning systems start analysing the training data to build their knowledge and produce their own rules and classifiers.

Machine learning techniques can be divided into three categories [14] [15]:

- supervised learning whose models are trained using the labelled data set, where the model learns about each type of data, that includes models like hidden Markov model (HMM) [16] and support vector machines (SVM) [17].
- semi-supervised learning that involves a small degree of supervision, and one example is bootstrapping [18].
- unsupervised learning whose model is not trained. The most common approach of the unsupervised category is clustering [19] [20].

Deep learning is a subfield of machine learning based on artificial neural networks which try to learn from the layered model of inputs. In the deep learning approach concept, learning of a current layer is dependent on the previous layer input. Deep learning algorithms can fall into both supervised and unsupervised categories [21]. The main applications of deep learning include pattern recognition and statistical classification. For example, to combat the increasing amount and reduce the threat of malicious programs, novel deep learning was developed [22].

In this work, we develop an aid system for medication intake — a chatbot — that gives answers to queries related to the medicine intake. Previously, the system had to been informed about the user's prescription, that is, the daily medicine intake routine and detailed information about every drug. We use the subtractive mountain clustering (SMC) algorithm to perform NLP tasks.

The main objective of this work is to apply the SMC algorithm within the chatbot to facilitate communication. Namely to process natural language using the SMC algorithm.

The topics covered in this paper relate mainly to chatbots and NLP. To conclude this section, we start by recalling some important facts in the history of chatbots in Subsection 1.1 and in Subsection 1.2 we review some existing chatbot solutions in healthcare. In Section 2, the chatbot configuration is described. Then, in Section 3, an overview of NLP and its implementation is discussed. In Subsection 3.1, we present the required steps for text pre-processing and, in Subsection 3.2, the

SMC algorithm, necessary to build the response, is described. Finally, in Section 4, we present and discuss the results obtained for the case study. We conclude with Section 5 where the main outcomes of the work are outlined and some guidelines for future work are stated.

### 1.1 History of chatbots

The first chatbot was created in 1966 by Joseph Weizenbaum. It was called ELIZA [23], and the objective was to pretend to be a psychologist. To do this, it used simple rules of conversation and rephrased most of what the users said to simulate a Rogerian therapist — person-centreed therapy. In 1991, the Loebner Prize, an annual competition in artificial intelligence, was launched. The contest awards the computer programs considered to be the most human-like. The competition takes the format of a standard Turing test, i.e., in each round, a human judge simultaneously holds textual conversations with a computer program and a human being via computer and, based upon the responses, the judge must decide which one is which [24]. In 2014, a chatbot named Eugene Goostman managed to fool 33% of the judges, thereby beating the test.

Another example of a chatbot is using Natural Language Interface to Database (NLIDB) to access information in the databases instead of Structured Query Language (SQL). An NLIDB system is proposed as a solution to the problem of accessing data in a simple way: any user can access the information contained in the database and get the answer in natural language [25]. Nowadays, multiple virtual assistants already exist, being the most complex and more widely used. Siri (from Apple), Google Assistant (from Google) and Alexa (from Amazon). At the moment, they are mainly used to call people, ask for directions or search the Internet for information [26].

### 1.2 Chatbots in healthcare

These days, chatbots also have increased use in healthcare to treat disorders such as cancer [27] or induce behavioural changes such as quitting smoking [28] and weight control [29]. Chatbots are increasingly being adopted to facilitate access to information from the patient side and reduce the load on the clinician side. In the field of medicine, there are already chatbots for the most varied purposes. Next, some illustrative examples are reviewed:

1. **OneRemission** is a healthcare chatbot to help cancer survivors, fighters, and supporters to learn about cancer and post-cancer health care [30].
2. **Wysa** is an emotionally intelligent chatbot. Its purpose is to help the user to build mental resilience and promote mental well-being with a test-based interface [31].

3. **Florence** is a chatbot related to medication intake that can remember taking medication, monitor certain biomedical parameters, and find information about diseases [32].

After reviewing the chatbots already developed in healthcare, we did not find one that focuses on taking medication by older adults, like the chatbot presented here.

## 2 Chatbot configuration

The chatbot herein presented was designed to help older adults with their medication. The chatbot can provide information about the physician's prescription, the medicine's package insert, and also extra information, such as the colour of the box and the colour of the pill, to avoid confusion in taking medication. Regarding the medical prescription, the chatbot can inform about the dose (how many pills per day) and when to take it (at which part of the day). Another set of information relates to the medicine's package insert, for example, indications, side effects, and what to do in case of forgetting to take it. Finally, the chatbot can also provide an image of the medicine box and the pill so that the patient can easily identify the medicine to be taken. Once the medicine in question is identified, then the related information can be retrieved.

The set of questions and answers must be as similar as possible to a conversation between human beings. Chatbots are developed to connect with the users and feel that they are communicating with a human and not a bot. In our chatbot, the possible answers are predefined and designed to emulate daily human communication.

Since a main step in the construction of the chatbot is NLP, this topic will be discussed in Section 3.

## 3 Natural language processing

NLP is an area of computer science, more specifically, in artificial intelligence (AI), concerned with giving computers the human ability to understand text and spoken words. This processing generally involves translating natural language into data (numbers) that a computer can use and generate a certain answer. NLP is applied in several areas, such as, machine translation [33], text summarisation [34] and spam detection [35]. One of the NLP's uses is chatbots.

A chatbot system analyses a query posed by a person and generates an answer from an organised collection of data stored and accessed electronically from a computer system. Usually, the answer is retrieved based on the basic keyword matching, and a selected response is then given as the output. When we talk about natural language, there are many ways to say the same information. So, when the chatbot is faced

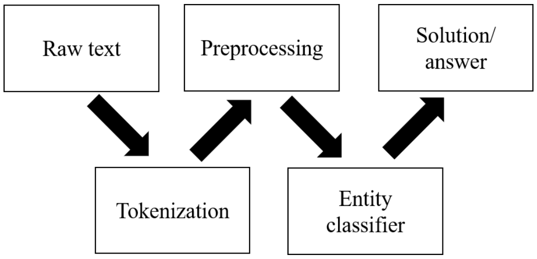


**Fig. 1.** Flowchart of the process needed from input to response in the chatbot.

with several alternative sentences requesting the same information, it is necessary to use an algorithm that can import what is truly relevant. After selecting this information, the chatbot will be able to insert the phrase into a context to produce the proper answer. Furthermore, the data needs to be pre-processed so that the chatbot can easily understand it.

In Figure 1, one can see a flowchart that shows the process necessary for the chatbot to generate an answer fitting the user input text. In the following subsection, we will detail the pre-processing of text.

We apply word processing to both (i) the drug's package insert, which is used to define the algorithm that leads to the answers, and (ii) the user's queries.

### 3.1 Text pre-processing

Natural Language Understanding (NLU) converts natural language utterances into a structure that the computer can deal with. As a sentence cannot be processed directly to the model, it needs to perform some NLP to further operations.

To achieve this, we implement an algorithm in MATLAB that uses a set of functions included in the *Text Analytics Toolbox.* Hence:

1. Read the file with raw text.
2. Segment the text into paragraphs — documents.
3. Remove empty documents.
4. Segment the text into tokens — tokenisation. This process consists in splitting the text into tokens, which are the basic units.
5. Remove unwanted words (stop words) or irrelevant punctuation from those tokens. Stop words are the words present in a sentence that does not make much difference if removed. For example, the words and, off, for.

6. Use normalising techniques, which consist in finding the root/stem of a word. For example, the words "ends" and "ending" are represented by the same stem - "end".

The various steps above are not always an easy task to carry out. Non-standard words are often ambiguous. Some words can have different meanings depending on the context. Moreover, there are also some acronyms and abbreviations that can be misleading. For example, should an acronym be read as a word (IKEA) or using each letter in a sequence (IBM)? The abbreviation "Dr." has a full stop that can lead to the wrong separation of sentences.

The following step is word embedding, which consists in converting text to numbers. Converting words into numbers will make the algorithm easier to apply [36] [37] [38].

### 3.2 Building the response

Creating an algorithm that could relate the user's queries to the answers that contain the right information is a big deal for chatbots.

Before selecting the correct answer is necessary to identify to which class/topic the query belongs. Here, the SMC algorithm is used to define word sets (clusters) [39].

Based on the medication package's insert is defined a proximity degree between the words related to medication intake. The calculus of the proximity degree is based on the distance between any two words of the package insert. Only relevant words are taken into account to calculate the distance. The word frequency and size are taken into account to evaluate relevance. One possible criterion is, for instance, a word appearing more than twice and having more than two letters being classified as relevant.

The calculus of the distance between two relevant words is based on their relative position in the text, where $r$ and $s$ are the codes of the words that occupy positions $i$ and $j$, respectively. Hence, we distinguish three different cases:

- Two words are in the same sentence: the distance reflects the number of words that separate them. The end of sentence recognition is done by punctuation marks. Then, $D(r, s) = |j - i|$.
- Two words are in the same document but in different sentences: the distance reflects the number of sentences that separate them and the number of words. Let $S_n$ be the number of separation sentences, then their pair distance is given by $D(r, s) = |j - i| \, S_n a$, where $a$ is an adjustable parameter according to the average number of words in each sentence.
- Two words are in different documents: the distance only reflects the number of documents that separate them. Let $P_n$ be the number of separation paragraphs,

then their pair distance is given by $D(r, s) = P_n b$, where $b$ is an adjustable parameter according to the number of words in sentences and sentences in the paragraph.

This is indeed a distance, since is non-negative, symmetric and verify the triangular inequality.

As the words may appear many times, the minimum distance between them is considered.

In addition, we also need a factor $B(r, s)$ for each word pair that reflects the number of times each pair appears together in the same sentence and/or the same paragraph.

In the SMC algorithm, we use the distance and $B(r, s)$ to calculate the potential. The algorithm is presented next.

**Subtractive mountain cluster algorithm** Sometimes, it is necessary to reduce the size of the data set to a set of representative points. For example, fuzzy logic algorithms are very complex and are not applicable in a large data set [40].

The SMC approach, developed by Chiu [39], assumes that enormous data sets are partitioned into subsets called clusters, and each cluster is represented by one representative element called the cluster centre. Initially, all data set points are potential cluster centres and each point potential is calculated by equation (1). Hence

$$P(I) = \sum_{j=1}^{N} e^{-\alpha |x_i - x_j|^2}, \tag{1}$$

for i = 1, . . . , N, ($N$ is the number of points) and $\alpha = 4/(r_a)^2$ for constant $r_a > 0$. Equation (1) shows that a data point with many neighbouring data points will have a high potential value. Parameter $r_a$ defines the data points influence. Data points outside this radius have little influence on potential.

The SCM algorithm allows for the association of a set of words into clusters. After finding the centre of every group — the most relevant term, all the others are aggregated according to a defined metric adapted to the language processing realm.

We adapted equation (1) to NLP and define equation (2). So the greater the number of times a word pair appears, the greater the potential of that word pair. Hence

$$P(r, s) = B(r, s) \sum_{s=1}^{N} e^{-\alpha D(r,s)^2}, \tag{2}$$

where $D$ is the symmetric distance matrix, $N$ the number of words. We assume the potential of equals words pairs is approximately zero.

After the potential of every distance has been computed, as the matrix $D$ is symmetric, we sum up all the potentials for each and every word, and we obtained a vector with $N$ columns. Thus, we were able to obtain the potential for every word and choose the one with the greatest potential as the centre of the first cluster. Let $I_1^*$ be the first word cluster centre index and $P_1^*$ its potential value.

Equation (3) is an adaptation of the potential equation for new cluster centre words after determining the first cluster centre word. Hence

$$P^*(r, s) = P(r, s) - P_1^* B(r, s) e^{-\beta D(I_1^*,s)^2}, \tag{3}$$

where $\beta = 4/(r_b)^2$ for constant $r_b > 0$. With Equation (3), we subtract a portion of potential at each word pair according to the distance from each word to the word chosen as the centre of the first cluster. Words close to the first cluster centre will have significantly reduced potential and are unlikely to be selected as the next cluster centre. Parameter $r_b$ defines the radius affected to the potential reduction.

Now, the word with more potential is selected as the second cluster centre. We then further reduce the potential of each word according to their distance to the second cluster centre. In general, each time we select the next centre of the next cluster, we revise the value of the potential in the following manner:

$$P_k^*(r, s) = P^*(r, s) - P_{k-1}^* B(r, s) e^{-\beta D(I_k^*,s)^2}. \tag{4}$$

This iterative process ends when the word with the most potential $P_k^*$ is less than $\epsilon P_1^*$ where $\epsilon$ is a small fraction.

The minimum distance $d_{min}$ between every two cluster centres also needs to be defined. If the following inequality is verified, the point is accepted as a possible centre of a cluster:

$$\frac{d_{min}}{r_a} + \frac{P_k^*}{P_1^*} \geq 1.$$

Next, the belonging degree of each word to the cluster is calculated using equation (5):

$$U(I) = \frac{1}{\sum_{k=1}^{c} \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \tag{5}$$

where $m$ is the hyper-parameter that controls how fuzzy the cluster is.

After applying the SMC algorithm, we can get groups of words (clusters) according to the distance they appear in the medicine package insert. It is necessary to pre-process the text of the questions, extracting the relevant information. Therefore, we can attribute them to a cluster and generate an appropriate answer.

## 4 Results and discussion

The parameters used are available in Table 1. We used the Xarelto drug package insert to validate the algorithm, which is used to prevent blood clots.

**Table 1.** Parameters

| ***Parameters*** | **Value** |
|---|---|
| $a$ | 10 |
| $b$ | 20 |
| $r_a$ | 12 |
| $r_b$ | 14 |
| $\epsilon$ | 0.1 |
| $m$ | 2 |

We obtained 297 documents (paragraphs) and 838 tokens (include words and two important punctuation marks).

Figure 2 shows the number of times the first 40 most frequent words appear. The most frequent word is the name of the drug — ”Xarelto” and then ”patient”. After the first ten words, the frequency of each expression remains approximately the same.

After applying the word relevance metrics, we get 468 relevant terms representing 55.85% of the total words.

The distance between relevant words is represented in Figure 3. On each axis, we have the indices of the words. The closer to white the colour, the smaller the distance between words.

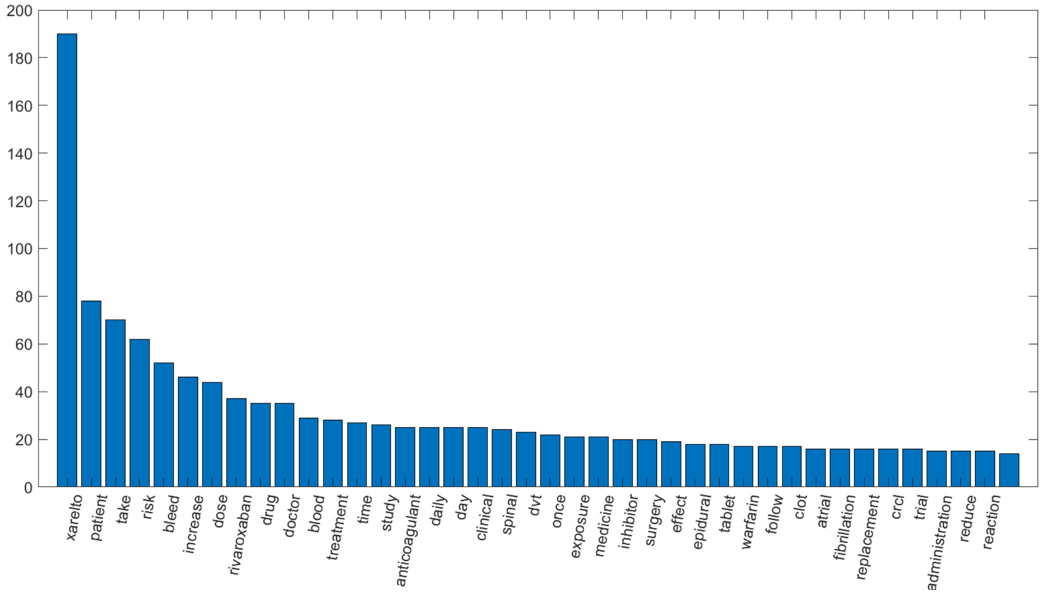


**Fig. 2.** Number of times the first 40 most frequent words appear.

We get 20 clusters and the first five are represented in Figure 4. In the graphs, the degree of belonging is also represented. We can see that in each cluster there are at least 14 words with a degree of belonging equal to 1.

In addition, we calculate the degree of belonging of each word to the cluster and in Figure 5 is the graph that shows the number of words belonging to each cluster with a degree of belonging greater than 0.5.

We can get the most significant words, with question pre-processing, and associate them to a cluster or set of clusters. The answer is in accordance with the belonging relationship of words to clusters.

For example, one question with the stemming words "risk", "foetal" and "bleed" has multiple possible answers. We can sort answers by the relative relevance. Answers correspond to the paragraphs (documents) whose profile of belonging to the words in the question and the document are closest. The answer with 1.00 as relative relevance is "Xarelto increases the risk of bleeding and can cause serious or fatal bleeding. In deciding whether to prescribe Xarelto to patients at increased risk of bleeding the risk of thrombotic events should be weighed against the risk of bleeding."

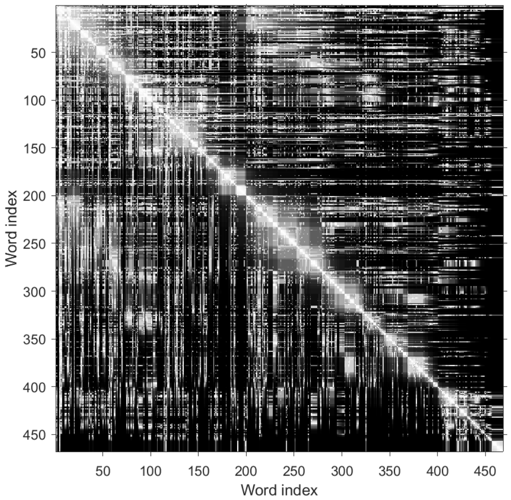


**Fig. 3.** Representative diagram of word distance. The white colour represents the closing words.

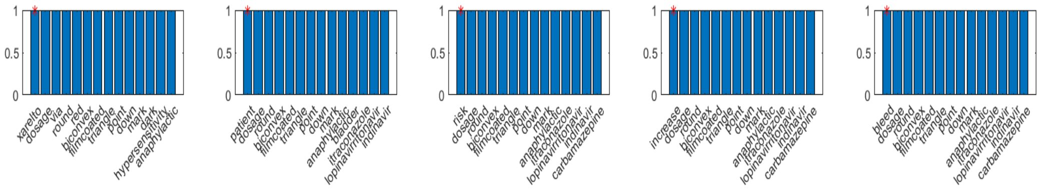


**Fig. 4.** First five word clusters in function of their normalising belonging degree.

## 5 Conclusion and future work

Chatbots are present in different areas, including healthcare. In this field, there are chatbots for many purposes, from psychologist chatbots to medication reminders.

Chatbots have the potential to help elderly people in taking medication since they can emulate human conversation. However, after a thorough literature review, we conclude that they are still not widely used to solve and avoid medication errors.

A fundamental step to enable communication between the users and chatbots is the processing of natural languages. The NLP area involves a large scientific community and contains a variety of associated algorithms.

We adapt the SMC algorithm to NLP. To the best of our knowledge, this algorithm has not been used before to resolve this type of problem. This algorithm does not define *a priori* the number of clusters, leaving rather more freedom to the association. Instead, through word clusters creation based on distance, we were

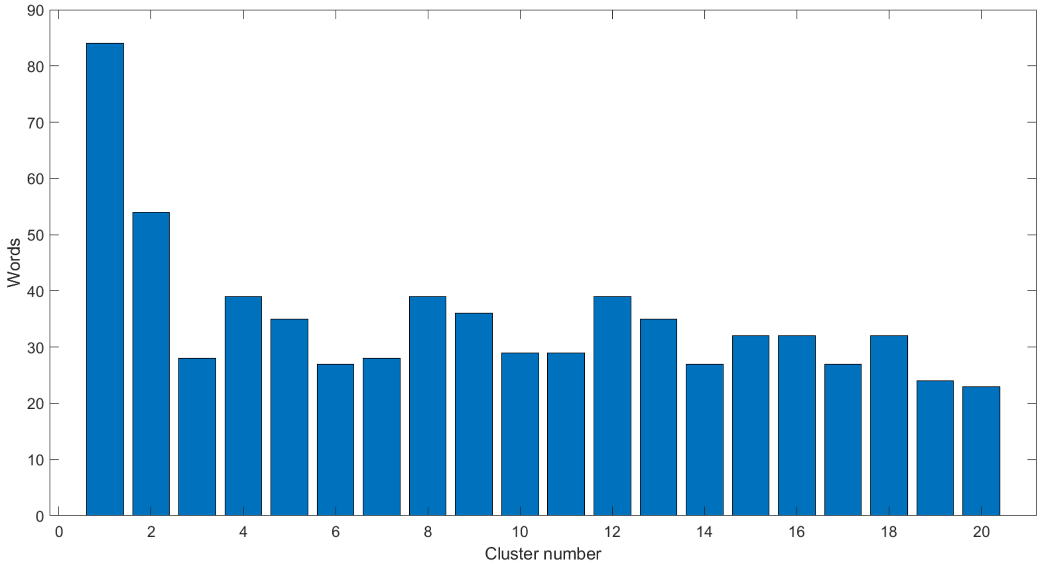


**Fig. 5.** The number of words belonging to each cluster with a degree of belonging greater than 0.5

able to define a degree of belonging among words and thus build adequate answers to the posed queries.

In this work, an aid system to help elderly people with medicine intake is proposed. Namely, a chatbot that is able to interact with the user in natural language. To process natural languages, a fundamental step is to achieve a communication system--user. To do this, we apply the SCM algorithm to define the relationship between words. Furthermore, we implement the proposed solution and applied it to a simple problem, obtaining results that we consider promising.

To complete this work, building a user interface might be the next step. The refinement of the algorithm is also in place. Moreover, the study of the accuracy of the SCM model should be done as well as to investigate whether other algorithms can solve the same problem. In addition, a study of the adherence of people of more advanced age to chatbots might be also an interesting point of investigation.

## References

1. E. D. Kantor, C. D. Rehm, J. S. Hass, A. T. Chan, and E. L. Giovannucci, “Trends in Prescription Drug Use Among Adults in the United States From 1999-2012,” *JAMA*, vol. 314, no. 17, pp. 1818–1830, 2015.
2. S. Reddy. (2017, Jul.) Patients Make More Medication Mistakes. [Online]. Available: https://www.wsj.com/articles/patients-make-more-medication-mistakes-1500913414

3. J. M. Ruscin and S. A. Linnebur. (2018, Dec.) Aging and Drugs. [Online]. Available: https://is.gd/ulrQx1
4. J. J. Mira, S. Lorenzo, M. Guilabert, I. Navarro, and V. Pérez-Jover, "A systematic review of patient medication error on self-administering medication at home," *Expert Opinion on Drug Safety*, vol. 14, no. 6, pp. 815–838, 2015.
5. A. Berman, "Reducing Medication Errors Through Naming, Labeling, and Packaging." *Journal of medical systems*, no. 28, pp. 9–29, 2004.
6. M. E. Reed, "The Health Information Technology Special Issue: New Real-World Evidence and Practical Lessons," *The American Journal of Managed Care*, vol. 25, no. 1, p. 12, 1 2019.
7. A. Ahmed, M. R. R. A., and Barua, "Home Medication for Elderly Sick People," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, 2019, pp. 369–374.
8. M.-Y. Wang, P. Tsai, J. Liu, and J. K. Zao, "Wedjat: A Mobile Phone Based Medicine In-take Reminder and Monitor," in *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*, 2009, pp. 423–430.
9. J. Zao, M. Wang, P. Tsai, and J. Liu., "Smart phone based medicine in-take scheduler, reminder and monitor," in *The 12th IEEE International Conference on e-Health Networking, Applications and Services*, 2010, pp. 162–168.
10. A. S. Gillis. (2020, Jan.) Bot (robot). [Online]. Available: https://whatis.techtarget.com/definition/bot-robot
11. M.-H. Wen, "A conversational user interface for supporting individual and group decision-making in stock investment activities," in *2018 IEEE International Conference on Applied System Invention (ICASI).* IEEE, April 2018, pp. 216–219.
12. T. Schmidlen, M. Schwartz, K. DiLoreto, H. L. Kirchner, and A. C. Sturm, "Patient assessment of chatbots for the scalable delivery of genetic counseling," *Journal of Genetic Counseling*, vol. 26, no. 6, pp. 1166–1177, 2019.
13. E. Sandalova, J. G. Ledford, M. Baskaran, and S. Dijkstra, "Translational Medicine in the Era of Social media: A Survey of Scientific and Clinical Communities," *Frontiers in Medicine*, vol. 6, p. 152, 2019.
14. S. S. Dash, S. K. Nayak, and D. Mishra, "A Review on Machine Learning Algorithms," in *Intelligent and Cloud Computing*, vol. 153, 2021, pp. 495–507.
15. W. Khan, A. D. J. Nasir, and T. Amjad, "A survey on machine learning models for Natural Language Processing (NLP)," *Kuwait Journal of Science*, vol. 43, pp. 95–113, 10 2016.
16. N. Ebrahimi, A. Trabelsi, S. Islam, A. Hamou-Lhadj, and K. Khanmohammadi, "An HMM-based approach for automatic detection and classification of duplicate bug reports," *Information and Software Technology*, vol. 113, pp. 98–109, 2019.
17. M. S. Desai, S. P. Gururaj, and T. L. Prakash, "Analysis of Health Care Data Using Natural Language Processing," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 242–246.
18. S. Deepika and T. Geetha, "Pattern-based bootstrapping framework for biomedical relation extraction," *Engineering Applications of Artificial Intelligence*, vol. 99, p. 104130, 2021.
19. A. Chokor, H. Naganathan, W. K. Chong, and M. Asmar, "Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning," *Procedia Engineering*, vol. 145, pp. 1588–1593, 2016.
20. G. Liu, M. Boyd, M. Yu, S. Z. Halim, and N. Quddus, "Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques," *Process Safety and Environmental Protection*, vol. 152, pp. 37–46, 2021.
21. S. Weidman, *Deep Learning From Scratch: Building With Python From First Principles*, 1st ed. O'Reilly Media, Incorporated, 2019.
22. J. Zhang, "CLEMENT: Machine Learning Methods for Malware Recognition Based on Semantic Behaviours," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, 2020, pp. 233–236.

23. J. Weizenbaum, “ELIZA — a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
24. S. Janarthanam, *Hand-On Chatbots and Conversational UI Development*, 1st ed. Packt Publishing Ltd., 2017.
25. A. Kumar and K. Vaisla, “Natural Language Interface to Databases: Development Techniques,” *Elixir Comp. Sci. and Engg.*, vol. 58, pp. 14 724–14 727, 05 2013.
26. L. Burbach, P. Halbach, N. Plettenberg, J. Nakayama, M. Ziefle, and A. C. Valdez, “”Hey, Siri”, ”Ok, Google”, ”Alexa”. Acceptance-Relevant Factors of V]irtual Voice-Assistants,” in *2019 IEEE International Professional Communication Conference (ProComm)*, 2019, pp. 101–111.
27. R. V. Belfin, A. J. Shobana, M. Manilal, A. A. Mathew, and B. Babu, “A Graph Based Chatbot for Cancer Patients,” in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, 2019, pp. 717–721.
28. D. Calvaresi, J. Calbimonte, F. Dubosson, A. Najjar, and M. Schumacher, “Social Network Chatbots for Smoking Cessation Agent and Multi Agent Frameworks,” 10 2019, p. 7.
29. C.Huang, M. Yang, C.Huang, Y. Chen, M. Wu, and K. Chen, “A Chatbot supported Smart Wireless Interactive Healthcare System for Weight Control and Health Promotion,” 12 2018, pp. 1791–1795.
30. Keenethics. (2021, Jul.) OneRemission. [Online]. Available: https://keenethics.com/project-one-remission
31. B. Inkster, S. Sarda, and V. Subramanian, “An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study,” *JMIR Mhealth Uhealth*, vol. 6, no. 11, p. e12106, Nov 2018.
32. M. Bates, “Health Care Chatbots Are Here to Help,” *IEEE Pulse*, vol. 10, pp. 12–14, 05 2019.
33. Z. Zong and C. Hong, “On Application of Natural Language Processing in Machine Translation,” in *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 2018, pp. 506–510.
34. S. Jugran, A. Kumar, B. S. Tyagi, and V. Anand, “Extractive Automatic Text Summarization using SpaCy in Python NLP,” in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 582–585.
35. P. Garg and N. Girdhar, “A Systematic Review on Spam Filtering Techniques based on Natural Language Processing Framework,” in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2021, pp. 30–35.
36. M. Virkar, V. Honmane, and S. U. Rao, “Humanizing the Chatbot with Semantics based Natural Language Generation,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 891–894.
37. P. C. Gaigole, L. H. Patil, and P. M. Chaudhari, “Preprocessing Techniques in Text Categorization,” *IJCA Proceedings on National Conference on Innovative Paradigms in Engineering & Technology 2013*, vol. NCIPET 2013, no. 3, pp. 1–3, December 2013.
38. D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. Prentice Hall PTR, 2000.
39. S. Chiu, “Fuzzy Model Identification Based on Cluster Estimation,” *Journal of the Intelligent and Fuzzy Systems*, vol. 2, pp. 267–278, 01 1994.
40. A. Chmielowiec, *Chapter 3. Implementation of the mountain clustering method and comments on its practical use for determining clusters centres*, 05 2020, pp. 45–56.

## Authors

**Neuza Claro** is a Biomedicine 3rd student at University of Trás-os-Montes e Alto Douro. She enjoys programming and being in nature.

**Paulo A. Salgado** received the B.S. degree and the M.S. degree in electrical and computer science from the Faculty of Engineering, University of Porto, Portugal, in 1989 and 1993, respectively, and the Ph.D. degree in electrical and computer science from UTAD University, Portugal, in 1999. He is currently an Associate Professor with the Department of Engineering, School of Science and Engineering of UTAD, Portugal. His research interests include artificial intelligence, control, and robotics science.

**T-P Azevedo Perdicoúlis** graduated in Mathematics (Computer Science) at the University of Coimbra in 1991, and pursued graduate studies at the University of Salford, Manchester (UK), where she obtained an MSc degree in Electronic Control Engineering in 1995 and a Ph.D. degree in Mathematics and Computer Science in 2000. She works for the Universidade of Trás-os-Montes e Alto Douro, Portugal, since 1991 where she is currently an Associate Professor. She is a founding member of the research institute ISR, University of Coimbra, and also of APCA, IFAC. Her main research interests are in the fields of differential games, optimisation and simulation of gas networks as well as identification theory. Besides research she has got an intense pedagogic activity, having taught many different courses in applied mathematics to engineers.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# TIPPING THE SCALES: A CORPUS-BASED RECONSTRUCTION OF ADJECTIVE SCALES IN THE MCGILL PAIN QUESTIONNAIRE

Miriam Stern

Program in Linguistics, Princeton University, Princeton, New Jersey, USA

## *ABSTRACT*

*Modern medical diagnosis relies on precise pain assessment tools in translating clinical information from patient to physician. The McGill Pain Questionnaire (MPQ) is a clinical pain assessment technique that utilizes 78 adjectives of different intensities in 20 categories to quantify a patient's pain. The questionnaire's efficacy depends on a predictable pattern of adjective use by patients experiencing pain. In this study, I recreate the MPQ's adjective intensity orderings using data gathered from patient forums and modern NLP techniques. I extract adjective intensity relationships by searching for key linguistic contexts, and then combine the relationship information to form robust adjective scales. Of 17 adjective relationships predicted by this research, 10 show agreement with the MPQ, which is statistically significant at the .5 alpha level. The results suggest predictable patterns of adjective use by people experiencing pain, but call into question the MPQ's categories for grouping adjectives.*

## *KEYWORDS*

*Corpus Construction, Adjective Scales, Pain Assessment, McGill Pain Questionnaire.*

## 1. INTRODUCTION

The question of pain's communicability is a crucial one; to treat pain, it must first be identified and categorized. In the past few decades, the practice of using numbers to describe pain has been questioned by linguists, medical professionals, and others [1],[2],[3]. Specifically, studies have shown that clinical data that takes only pain intensity into consideration is insufficient [1],[3]. This paper considers the McGill Pain Questionnaire (MPQ), the most commonly used verbal pain assessment tool that asks patients to describe their pain using a provided list of adjectives [1],[4]. In collecting a combination of adjectives, the MPQ is meant to extract more nuanced information than can a numerical rating system that considers only intensity data [1][2].

The MPQ has several sections, but this research focuses on the one entitled "What Does Your Pain Feel Like?" This section asks patients to select up to one word in each of 20 adjective categories that describes their pain. Each category contains between two and six words, which the MPQ posits are gradations of the same sensation, and which are assigned a numerical value accordingly. For instance, category 11's "tiring" and "exhausting" are different intensities of one scalar property, with "tiring" assigned a value of 1 and "exhausting" a value of 2. In this way, the MPQ seeks to translate quantitative verbal descriptions into qualitative numerical data to communicate pain [1].

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 259-275, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111421

An analysis of the body of literature on pain assessment tools reveals that various components of the MPQ have been independently retested by other researchers since its construction [5]-[8]. Several studies have shown that the MPQ generally has good construct validity [6]-[8], however, one study called into question the correctness of the MPQ's adjective groupings, suggesting that the adjectives might be imprecisely categorized [5]. There are additional weaknesses revealed in the studies conducted to construct and then verify the MPQ. Firstly, the original MPQ research used mixed populations of doctors and patients to construct the adjective intensity scales, with the two groups being observed to assign different values to pain adjectives [2]. Since the goal of pain assessment questionnaires is to facilitate the communication of a patient's pain, the ways in which patient's actually communicate about their pain is vitally important. Additionally, all studies conducted on the MPQ's efficacy have been laboratory experiments, with the adjectives provided to the patients by the researchers. Thus, the patients' interactions with the adjectives were not entirely natural [5]-[8]. Finally, since the MPQ was constructed in the 1970's, and language is constantly evolving, the time-dependency of the MPQ's adjectives must also be called into question.

My research improves upon prior methods by analyzing corpus data drawn from self-authored forum postings of pain sufferers. To assess the MPQ's system of ranking adjectives by intensity, I created a text corpus from online chronic pain forum posts. Using this corpus, I conducted a search for specific linguistic contexts in which adjectives are used; from these contexts, the intensity relationships amongst pairs of adjectives was inferred. By combining these generated intensity relationships, I constructed novel adjective intensity scales and compared them to the ones present in the MPQ. This approach allows for the analysis of spontaneously produced pain descriptions, which reflect each author's pain most personally. Furthermore, the contemporariness of the blog postings provided insights into the timelessness or lack thereof of the adjective scales devised by the MPQ's creators Finally, the construction of these adjective scales requires that patients use adjectives in a consistent manner to describe their pain. As such, successful reconstruction of adjective intensity scales would validate the concept behind using an adjective questionnaire to elicit medical and diagnostic data.

As a preliminary study, this data corroborates the adjective intensity scales defined by the MPQ at above chance levels. This research suggests, however, that the categories for the MPQ adjectives may be imprecisely defined. Accordingly, further research will be required to fully analyze the validity of the MPQ.

## 2. LINGUISTIC BACKGROUND

To understand how verbal pain questionnaires work, a discussion of adjective scales in general is needed. In this section, I will introduce the concept of "scalar implicature", and then consider its role in adjective meaning and intensity.

### 2.1. Introduction to Scalar Implicature

There is a distinction drawn in linguistics between meaning that is conveyed explicitly or verbally and meaning that is implicitly derived. For example, consider the following statement:

a. Ezekiel likes some of the teachers in his school.

At surface level, there is a literal meaning conveyed by the words of this statement. However, in any world in which (1a) is true, (1b) must also be true. This is an example of *entailment*, which is meaning conveyed automatically and immutably by a statement. Additionally, someone who heard statement (1a) would typically understand it to imply (1c) [9].

b. Ezekiel likes at least one teacher in his school.
c. Ezekiel does not like all of the teachers in his school.

The relationship between (1a) and (1c) is an example of *implicature*, which refers to meaning that is implied by an utterance without being explicitly stated. Unlike (1b), the statement made by (1c) can be cancelled, as demonstrated by (1d), or reinforced, as demonstrated by (1e) [9],[10].

d. Ezekiel likes some of the teachers in his school; in fact, he likes them all.
e. Ezekiel likes some of the teachers in his school, but not all of them.

Though there are different types of implicature, this example demonstrates *scalar implicature*: involving words like 'some' and 'all,' scalar implicature is concerned with the ways in which word meanings differ in intensity, and the additional information that they can convey [9],[10]. In this case, conversational conventions dictate that if the speaker knew that Ezekiel likes *all* his teachers, the speaker would be expected to say that Ezekiel likes *all* of his teachers (since *all* is a more informative statement than *some*). The fact that the speaker said *some* and not *all* therefore implies that the speaker was not able to say *all* [11].

This concept of scalar implicature applies to other adjectives as well, which can be organized in "Horn scales" [9],[10]. As an example, on the scale <pretty, beautiful> *beautiful* entails *at least pretty*, though *pretty* does not entail *at least beautiful*. Thus, a term on the left of a pair like <pretty, beautiful> suggests that any term to the right is inapplicable, or at least not known to be applicable [10],[12].

These properties of scalar implicature and adjective scales form the basis of adjectival pain assessment tools. In particular, adjectival pain assessment tools rely on the assumption that, if a patient describes her pain using adjective A from among a particular scale, the patient is describing her pain as A, *to the exclusion of any stronger description* in the same category [1].

## 3. HYPOTHESIS

Building upon the body of literature surrounding adjective scales and pain questionnaires, I evaluated the adjective groupings presented by the MPQ. Specifically, I attempted to assemble Horn scales using the adjectives found in the MPQ by analyzing online forum data. Comparing these constructed scales to the categorical hierarchies prescribed by the MPQ provided insight into the ways in which people use adjectives to describe their pain, and thus into the MPQ's validity. This research sought to reject the null hypothesis $H0_a$ and affirm the hypothesis H1.

$H0_a$: There is no predictable pattern to the way in which people use scalable adjectives.
H1: There is a predictable pattern to the way in which people use scalable adjectives.

Additionally, using unique chronic pain forums dedicated to specific types of pain, this research considered whether people suffering from different diseases or ailments tend to use unique frequency distributions of adjectives to describe their pain. This leads to another set of hypotheses:

$H0_b$: Different categories of chronic pain are not associated with specific adjective pain descriptors.

H2: Different categories of chronic pain are associated with specific adjective pain descriptors.

The efficacy of the MPQ and other similar metrics depends on the predictable and unambiguous way in which scalable adjectives are utilized; an adjective description of pain is only useful insofar as its meaning is mutually agreed upon by the patient providing the adjective and the physician receiving it. Therefore, affirmation or rejection of these hypotheses provides useful insights into the rationality of using an adjectival pain scale for clinical diagnoses.

## 4. CORPUS

### 4.1. Web Scraping and Pre-Processing

For the textual data needed in this research, I created a corpus of text produced by patients experiencing pain. I chose internet forums as the source of this data because of their public and voluntary nature, and the wide range and specificity of forum topics. Crucially for this research, forum postings reflect speech occurring naturally amongst patients, rather than between patients and physicians.

For this paper, data was pulled from a website called HealingWell.com, which is meant to be an online community for those experiencing chronic pain [13]. Within this website, data was pulled from forums of the following three topics: Chronic Pain, Rheumatoid Arthritis (RA), and Fibromyalgia. Chronic Pain was chosen as a catch-all for the different types of pain that people experience. RA and Fibromyalgia were both chosen as conditions for which pain is a primary symptom, according to their Mayo Clinic descriptions [14],[15]. I selected these three categories to consider whether there might be a different distribution of adjective use in pain descriptions between the different categories—RA and Fibromyalgia—and between the two categories and the broader Chronic Pain forum.

From the HealingWell website, I scraped data from the three forums of interest [13],[16]. Specifically, I collected all the text from blog postings on each of the three topics. Then, I preprocessed the data by tokenizing the text [17], and correcting typos [18].

### 4.2. Corpus Description

Among the three forums (Chronic Pain, RA, and Fibromyalgia), the Chronic Pain forum was the largest, with a total of 20,189,291 words. Rheumatoid Arthritis was the second largest, with 4,160,952 words, and the smallest of the three was the Fibromyalgia forum, with 4,156,802 words. Since adjectives are of interest in this paper, I calculated additional statistics for the adjectives in each forum text. Basic statistics about the forum texts are summarized in Table 1.

Table 1. Forum Data Descriptions.

| | Chronic Pain | Rheumatoid Arthritis | Fibromyalgia |
|---|---|---|---|
| Total Words | 20,189,291 | 4,160,952 | 4,156,802 |
| Mean Post Length (words) | 145 | 123 | 117 |
| Median Post Length (words) | 101 | 89 | 75 |
| Post Length Range | 1, 3681 | 1, 3456 | 1, 5047 |
| Unique Tokens | 193,504 | 43,733 | 73,561 |
| Type/Token Ratio (%) | 1.770 | 1.051 | 0.9584 |
| Total Adjectives | 343,665 | 331,462 | 1,959,734 |
| Unique Adjectives | 14,486 | 14,603 | 34,196 |

## 4.3. Linguistic Processing

Once the initial data collection and processing was complete, I categorized and tagged the textual data using an off-the-shelf part-of-speech (POS) tagger [17]. An abridged list of parts of speech with their corresponding tag abbreviation is given in Table 2 [17].

Since the primary part of speech of interest in this paper is the adjective ('JJ'), I gave special care to these tags to ensure their proper identification. An example sentence from the data is given below in the form 'word/TAG'.

(1) 'the/DT pain/NN is/VBZ a/DT different/JJ pain/NN ,/, the/DT best/JJS way/NN i/NN can/MD describe/VB it/PRP is/VBZ a/DT deep/JJ burning/NN pain/NN ,/, not/RB a/DT neuropathy/JJ pain/NN either/DT ./.'

As can be seen from the example sentence, the default NLTK POS-tagger is not 100% accurate: here, the word 'burning' would be more accurately categorized as an adjective. I will address this problem in the next section.

Table 2. Select Part of Speech Tagset.

| Tag | Part of Speech |
|---|---|
| CC | Coordinating conjunction |
| DT | Determiner |
| IN | Preposition/subordinating conjunction |
| JJ | Adjective |
| JJR | Adjective, comparative |
| JJS | Adjective, superlative |
| MD | Modal could, will |
| NN | Noun, singular |
| NNS | Noun, plural |
| RB | Adverb |
| VB | Verb |

### 4.3.1. Special Considerations

In English, certain suffixes are associated with specific types of speech. For example, if a word is tagged with the suffix [-s], as in the word 'cat-s', that word can be assumed to be a noun, since nouns take the plural suffix [-s]. However, English has many ambiguous morphemes, or meaning-carrying units, which undermine generalizations like the one introduced for the [-s] suffix. For example, words ending with [-ing] might be categorized as a verb, noun, or adjective. Examples of this are given in sentences (2)-(4), using the English root 'burn.'

(2) The ceremonial *burning* of the torch takes place tomorrow.
(3) The *burning* building could be seen from miles away.
(4) A light was *burning* in the hallway.

'Burning' assumes the role of noun (2), adjective (3), and verb (4) in these three sentences, with its meaning parsed based on the syntactic and semantic position of the word in the sentence. For this research, it was especially important that adjectives like 'burning' not be categorized as verbs. As such, I added an additional condition to the pos-tagger, such that all words ending in '-ing' that immediately preceded a noun were tagged as adjectives. This was meant to ensure that a usage such as (3) would be appropriately labeled as ADJ, whereas 'burning' in (2) and (4) would be unaffected.

In the final stages of data analysis, I still found instances where MPQ adjectives were mislabeled as nouns specifically. To guarantee that all instances of MPQ adjectives would be collected, the final step in tailoring the POS-tagger consisted of feeding the tagger a list of all the adjectives from the MPQ; any time the adjective appeared in the right morphological form, the automatic tag was overridden with the adjective tag.

### 4.4. Corpus Comparison

I assembled the corpus used for this research from specifically oriented text: all the forum postings were on the topic of chronic pain. To understand how broadly applicable the results of this research are, it is important to consider how the makeup of this corpus compares to other more topic-neutral corpora. I selected two other corpora for this comparison. The Brown Corpus is a roughly million-word corpus compiled from different genres of texts published in English in 1961 [19]. Though the Brown corpus is small and somewhat outdated, I opted to include it in the comparison since it was released close to the date of the MPQ's creation; if there is any time-dependence on the frequency of adjective use, the Brown corpus might show divergence from the other corpora.

I also included for comparison the much larger, billion-word Corpus of Contemporary American English (COCA). Comprised of spoken and written text collected from 1990 through the present day, this corpus is meant to serve as a big-picture sample of American English [20]. I collected frequency data of each MPQ adjective in each of the three corpora.

After assembling the frequency counts of each MPQ adjective in each of the corpora, I calculated a cosine similarity rating for each pair. The cosine similarity ratings between the corpora were as follows: HealingWell and Brown: 0.794; HealingWell and COCA: 0.620; Brown and COCA: 0.776. The cosine similarity value ranges between 0 and 1, with 0 indicating no overlap between the two sets, and 1 indicating complete agreement [21]. Since the three cosine similarity ratings are all high, there does not appear to be significant or categorical disparities in MPQ adjective frequency across different corpora or different contexts.

A similar analysis was done within the HealingWell corpus itself in relation to the different forum topics considered. For each of the forum topics—RA, Fibromyalgia, and Chronic Pain—the frequency of the MPQ adjectives was calculated. Graphing the data, the overall patterning of adjective frequency was consistent across forum topics. RA appears to be somewhat of an outlier for a few adjectives, including 'sore', 'itchy', and 'aching', for which the RA frequency is significantly higher than the other two topics.

## 5. METHODOLOGY

### 5.1. Identifying Adjective Contexts

After creating a corpus, I conducted searches to find the adjective contexts needed to construct adjective intensity scales. This section outlines the process for defining and locating the linguistic contexts of interest.

The target for this search was the type of construction described by Horn, whereby the pattern of a sentence containing two adjectives can convey the intensity relationship between the two. Examples of Horn's adjective constructions are provided in (5)-(7) where adjective Y has a stronger intensity than adjective X [10]. For each construction form, an example sentence is provided to its right.

(5) X but not Y // Warm but not hot
(6) Not only X, but Y // Not only ugly, but grotesque
(7) X if not Y // Bright if not blinding

In addition to these examples from Horn, I also gathered intensity patterns by reviewing Sheinman et al. I utilized Sheinman's strategy of breaking down the examples into two categories: intense patterns and mild patterns. Intense patterns contain two adjectives X and Y such that Y is more intense than X [22]. An example of an intense pattern is given in (8).

(8) X, perhaps even Y // Good, perhaps even great

Mild patterns contain two adjectives X and Y such that X is more intense than Y. An example of an intense pattern is given in (9).

(9) Not Y but still very X // Not ginormous but still very big

I created a new list of each construction type using examples from Horn, Sheinman et al., and several novel patterns. The final list of intense and mild patterns totaled 13, with 6 intense and 7 mild patterns. I lay out the patterns of each type in Table 3.

I analyzed the constructed corpus data in 10-word chunks to search for matches against any of the specified mild and intense patterns. Though the intense and mild patterns are all 6 words or fewer, I opted to use n-grams with n=10 and to expand the regular expressions to allow for additional complexity in the phrases matched. That is, while the sentence 'all day at work it is **stiff but not painfu**l' would return a match using an n-gram with n=5, the phrase 'the pain is quite **sharp but not** usually particularly **achy**' would not. Using my approach, both phrases returned matches.

To further refine the search, I also ran the matched phrases against the list of MPQ adjectives. Phrases which contained two or more adjectives (at least one on either side of the intense/mild construction) from the MPQ list were selected, while all others were excluded. After this stage, I collected a total of 114 phrases: 27 from RA; 38 from Fibromyalgia; and 49 from Chronic Pain.

Table 3. Intense and Mild Patterns

| Intense Pattern | Mild Pattern |
|---|---|
| if not X at least Y | X but not Y |
| not X but Y enough | X but never Y |
| not X just/only Y | X but hardly Y |
| not X but still (very) Y | X even/perhaps Y |
| not/no X just/only Y | X perhaps/and even Y |
| no X just Y | X almost/if not/sometimes Y |
|  | X sometimes almost/even Y |

### 5.1.1. Excluded Data

Some phrases that matched the listed criteria were not suitable for analysis in this research. From the original 114 matching phrases, I discarded an additional 46 as 'false positives.' I excluded these phrases due to three remaining issues.

***Wrong Topic***. First, not all the identified phrases were on-topic; that is, adjective constructions that fit all the other criteria were sometimes describing non-pain-related subjects. Examples of phrases that were discarded for this reason are given in (10) and (11).

(10) you/prp melt/vbp the/dt wax/nn so/in its/prp$ **hot/jj but/cc not/rb burning/jj**
(11) i/prp love/vbp swimming/vbg but/cc **cold/jj** water/nn or/cc **even/rb cool/jj** water/n

Off-topic phrases represented the largest subset of rejected matches, with a total of 35 off-topic phrases discarded between the three forums.

***Wrong Tag.*** Secondly, as mentioned previously, POS-taggers often have difficulty distinguishing between gerunds (‘VBZ’) and modifiers (‘JJ’). To ensure that all MPQ adjectives were accounted for, the data processing included overriding all tags of MPQ adjective roots and replacing them with the adjective tag ‘JJ’. This default adjective tagging favored false positives over false negatives. As such, certain words suffixed with [-ing] were tagged as adjectives, even while being used as verbs or others in their contexts. Examples of phrases that were excluded for having mis-tagged adjectives are given in (12) and (13).

(12) **tight/jj** and/cc my/prp$ heart/nn **sometimes/rb** feels/nns like/in it/prp is/vbz **beating/jj**
(13) **tingling/jj** as/in **perhaps/rb** it/prp could/md be/vb **pressing/jj** on/in a/dt nerve/nn

Of 6 total phrases discarded for mis-tagging, four different words were mis-tagged: ‘killing’ (2), ‘beating’ (2), ‘pressing’ (1), and ‘cutting’ (1).

***Wrong Noun.*** Finally, some strings that matched the specified patterns contained multiple phrases in one sentence. Examples of this are given in (14) and (15).

(14) so/rb **annoying/jj** yes/uh **itchy/jj sometimes/rb** a/dt **hot/jj** soak/nn helps/vbz ./.
(15) not/rb suggest/vb just/rb using/vbg **cold/jj** pools/nns or/cc **even/rb hot/jj** tubs/nns

In (14), the adjectives ‘annoying’ and ‘itchy’ describe a sensation, whereas ‘hot’ modifies the noun ‘soak.’ Here, the issue is one of missing punctuation, which is common in internet writing. However, in (15), there are two clauses within one sentence separated by a coordinating conjunction, such that adjectives on either side of the conjunction modify two separate nouns. Since the two adjectives are being used separately and cannot readily be ranked on one intensity scale, an intensity relationship cannot be inferred. There were 5 total phrases rejected for including separately describing adjectives.

After I sorted through all the data and removed false positives, a total of 66 matched phrases remained for use in data analysis.

## 5.2. Adjective Scale Construction

Once I compiled and searched the corpus, I then used the data collected to address the research question pursued in this paper. In this section, I present the process of using the partially ordered weak-strong pairs discussed in Section 5.1 to construct adjective intensity scales.

### 5.2.1. Weak-Strong Pairs

Using the final list of matched phrases, each phrase was analyzed to yield one, or more, weak-strong adjective pair. First, I divided the phrases by the type of pattern they had matched: mild or intense. Within each group, I identified the conjunction string for each phrase based on the pattern matched. As an example, if a phrase matched the mild pattern ‘x and even y,’ the conjunction string would be ‘and even.’ For all mild patterns, I assigned the adjective on the left of the conjunction string the value *weak*, and the adjective on the right the value *strong*. If instead a phrase matched an intense pattern, I assigned the left adjective the value *strong* and the right adjective the value *weak*.

Where a phrase contained multiple adjectives on either side of the conjunction string, the same process was applied for each of the adjectives on either side of the construction, as shown using the example in (16).

(16) back/nn areas/nns can/md feel/vb **heavy/jj aching/jj** and/cc yes/jj <u>sometimes</u>/rb **burning/jj**

Here, there would be two strong-weak pairs identified—heavy-burning, and aching-burning—though there is no limit on the number of pairs that could be derived from one phrase.

In all, I identified 81 weak-strong pairs. Of the 81, 17 were found in the RA text, 25 in Fibromyalgia, and 39 in Chronic Pain. I then ran this list of pairs through a lemmatizer [17]. For the data at hand, this meant that instances of ‘itchy’ and ‘itching’ were identified with the same root of ‘itch.’ Similarly, ‘achy’ and ‘aching’ were identified with the same root of ‘ache.’ Though there are nuanced differences in meaning between the separately inflected forms of the same root, for the purposes of this preliminary analysis, this treatment of different inflectional forms was sufficient. After lemmatization, 27 of the MPQ’s 78 total adjectives were accounted for in at least one of the weak-strong pairs. The adjectives occurring in the greatest number of pairs were ‘burning’ (20), ‘sharp’ (19), ‘tingling’ (14), and ‘aching’ (13).

### 5.2.2. Adjective Categorization

Once I identified the weak-strong pairs, I then combined the partially ordered pairs to make categorical adjectival scales.

As mentioned previously, the MPQ divides its adjectives into 20 different categories. This organizational structure depends on the idea that, within each category, the adjectives all describe the same sort of sensation, differing only in intensity [1],[2]. Since our research’s list of 81 weak-strong pairs was derived from spontaneous speech, the two adjectives in each pair were not necessarily found in the same MPQ category. For example, consider the following sentence in example (17).

(17) no **hurting** or **prickling** feelings just the **numb tingling** feeling

In this sentence, there were four weak-strong pairs identified which involve adjectives from four different MPQ categories. The pairs, given in the form (weak: category, strong: category) are summarized as follows:

(18) numb:18, hurting:9
numb:18, prickling:3
tingling:8, hurting:9

tingling:8, prickling:3

Though there are clearly some differences in intensity between the adjectives, as demonstrated by the intense pattern 'no x, just y', the MPQ does not attempt to put adjectives of different categories on the same scale. This can be illustrated more clearly using adjectives that are not exclusively used to describe pain: cold/freezing and hot/burning. Though it is intuitively clear that 'freezing' is more intense than 'hot', and 'burning' more intense than 'cold,' the elements cannot all be incorporated into a Horn scale that maintains the principles of implicature [9],[10]. An example of a possible (19) and impossible (20) scale are given in the following examples:

(19) <cool, cold>
(20) *<cold, cool, warm, hot>

Using Horn's concepts of scalar implicature, if the scale given in (20) were possible, we would expect *cool* to implicate *not warm* just as *warm* implicates *not hot* [10]. Given the question "How is the temperature of your tea?" we would expect both (21) and (22) to be acceptable answers, where the implicature is expressed explicitly.

(21) It's warm, but not hot.
(22) *It's cool, but not warm.

Given that (22) is an unacceptable sentence, the scale provided in (20) is an impossible one; hence, I limited the analysis for this research to adjectives within the same MPQ category. However, since not all weak-strong pairs identified in the data involved adjectives of the same category, the next step in the process took advantage of transitive relationships between adjectives, as I will describe in the next section.

### 5.3. Graph Creation and Traversal

Constructing robust scales from the weak-strong pairs required a two-step process. First, I plotted all the weak-strong pairs involving MPQ adjectives on a graph. To construct the graph, a node was created for each adjective represented in the MPQ weak-strong pairs. For each pair, an edge was drawn between the two adjectives' nodes, with an arrow pointing from the weaker adjective to the stronger adjective [23],[24]. This graph is presented in Figure 1.

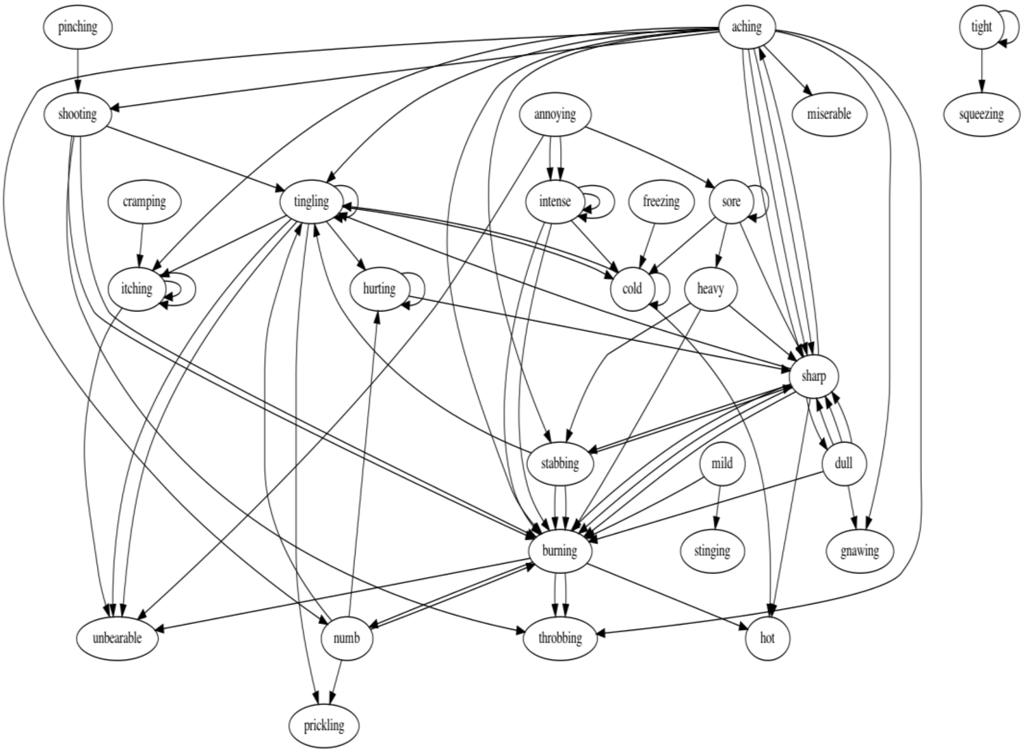


Figure 1. Weak-Strong Adjective Relation Graph

Where multiple arrows begin or end at a particular adjective node, that adjective was present in multiple weak-strong pairs. Several adjective pairs have bidirectional arrows between them. For example, in the case of ‘sharp’ and ‘dull,’ there are three arrows pointing from ‘dull’ to ‘sharp,’ suggesting that ‘sharp’ is stronger than ‘dull,’ and one arrow pointing from ‘sharp’ to ‘dull,’ suggesting that ‘dull’ is stronger than ‘sharp.’ These arrows are contradictory, as ‘sharp’ cannot be both stronger and weaker than ‘dull.’ Intuitively, ‘sharp’ is in fact stronger than ‘dull,’ which is reflected in the majority of the weak-strong pairs. In these cases, where arrows connecting two adjective nodes pointed in both directions, I prioritized the majority direction. I will explore this phenomenon, along with the rest of the graph, in more detail in the following data analysis section.

The second step of the scale construction process required graph traversal to find paths between adjectives of the same MPQ category. Ideally, each adjective in each MPQ category would have been found in a mild or intense pattern with another adjective of that same category. However, in the mined dataset, there were only eight instances where two adjectives of the same MPQ category were found in the same weak-strong pair. I present an example of this in (23), where both ‘annoying’ and ‘unbearable’ are category 16 adjectives.

(23) its an annoying pain but not unbearable pain its been

Because of this limited pool of data, I also considered transitive relationships between adjectives of the same MPQ category in the graph traversal. That is, given the two category 9 adjectives ‘aching’ and ‘hurting’, if ‘aching’ is weaker than ‘tingling’ and ‘tingling’ is weaker than ‘hurting,’ then ‘aching’ must also be weaker than ‘hurting.’ With transitivity, I was able to deduce intensity relationships even where there was not a direct connection between two adjectives of the same category. I traversed the graph using a recursive search of the graph’s nodes, given an input of every possible adjective pair combination within each MPQ category.

Starting at the node of one of the adjectives, each path away from the adjective was travelled from node to node until either a) there were no available and unvisited nodes to visit from the current node or b) an adjective of the same MPQ category was found. Using this recursive search, 404 total paths were traced between adjectives of the same MPQ categories, with 22 unique adjective pairs connected.

## 5.4. Scale Construction

After collecting the paths between adjectives and constructing the weak-strong pairs within each MPQ category, I combined the partial orderings to make robust adjective scales for each category. I considered each MPQ category independently. For each weak-strong pair in a given category, I assigned the weak adjective a value of '0' and the strong adjective a value of '1.' Then, I summed together the values for each adjective in a given category. I present an example of this process using category 16 adjectives (24).

(24) Weak-Strong Pair: <annoying, intense>

1. Annoying + 0
2. Intense + 1

Weak-Strong Pair: <annoying, unbearable>

1. Annoying + 0
2. Unbearable: +1

Weak-Strong Pair: <intense, unbearable>

1. Intense: +0
2. Unbearable: +1

Total Values:

Annoying: 0
Intense: 1
Unbearable: 2

Once I calculated the total value for each adjective, I constructed a scale that ranked the intensity relationships of the adjectives in each category in numerical order. For the example in (24), the category 16 adjectives would be ordered on the scale <annoying, intense, unbearable>, in ascending order of intensity.

The final constructed adjective scales for each of the categories represented in the data are presented in Table 4 below. Where two adjectives are separated by a slash, the data was inconclusive on the intensity relation between the two.

Table 4. MPQ Category Scale Reconstruction.

| MPQ Category | Adjective Scale |
|---|---|
| 3 | <stabbing, prickling> |
| 7 | <burning, hot> |
| 8 | <tingling, itching> |
| 9 | <dull/sore/aching, heavy, hurting> |
| 16 | <annoying, intense, unbearable> |
| 18 | <tight, squeezing> |
| 19 | <freezing, cold> |

# 6. DATA ANALYSIS

In this section, I analyze the data collected and discussed in Section 5, specifically as it pertains to the research questions introduced at the beginning of this paper. By exploring patterns in the weak-strong adjective frequency and elements of the adjective relation graph (Figure 1), I will examine the hypothesis that there is a predictable pattern to scalable adjective usage.

## 6.1. Weak-Strong Adjective Frequency

As mentioned in Section 5.3 only eight of 81 identified weak-strong pairs contained two adjectives from the same MPQ category. Of the remaining 73 pairs, many seem to be strongly related, though they are not grouped in the MPQ. Firstly, there are nine pairs of adjectives that are represented more than once in the weak-strong list. Some, like <intense, burning> are found twice in the same configuration, while others can be found ordered in both directions (See 5.3 for example with ‘sharp’ and ‘dull’). In either case, there seems to be some sort of connection between the adjectives such that they are more likely to be used together when describing a painful sensation. Beyond these nine pairs, there are other identified weak-strong pairs with specific relationships between adjectives that are not in the same MPQ category. Specifically, two weak-strong pairs— [sharp, dull] and [cold, hot] —are direct antonyms of one another (as defined by intuition and WordNet [25]). As discussed previously, it is impossible to order antonyms like *cold* and *hot* on a Horn scale such that scalar implicature applies [9],[10]. However, based on the data collected for this research, people sometimes use antonyms with the same constructions that otherwise suggest different gradations of meaning.

Finally, I identified additional weak-strong pairs which are not related by antonymy, but whose meanings are closely related or even synonymous. These pairs include examples like [tingling, prickling] and [aching, throbbing], which are not grouped together in the same MPQ category, but which are intuitively similar. To further explore the different types of relationships between adjectives found in the weak-strong pairs and in the MPQ in general, more robust comparisons using WordNet or survey data would be needed. However, the prevalence of adjectives from different MPQ categories grouped together suggests a need to reevaluate the division of the MPQ categories.

## 6.2. Adjective Graph

The graph labeled Figure 1 (Section 5.3) is a visual representation of the data collected in this research. In this section, I will explore a few key insights from the graph.

***Bidirectional Arrows***. As mentioned in the previous section, there are several adjective node pairs—five in total—which have arrows pointing both from adjective A to adjective B, and from B to A. These bidirectional node pairs show support for the null hypothesis: if an adjective A can be used in both a stronger and weaker position relative to adjective B, that would suggest that the relative intensity of adjectives is not predictable. However, with 22 adjective node pairs, the four bidirectional ones comprise only 23% of the total; the other 77% of adjectives were consistently ordered in relation to their connecting nodes. Interestingly, all the adjectives represented in bidirectional pairings—‘numb’, ‘burning’, ‘sharp’, ‘aching’, ‘dull’, ‘cold’, ‘tingling’—are in the top 40% of most frequently occurring MPQ adjectives within weak-strong pairs, and in the top 30% of the corpus overall. Further research could consider whether the frequency of adjectives has an impact on the consistency of usage, and specifically on how people perceive the intensity of high-frequency adjectives.

***Loops.*** In addition to bidirectional arrows, there are also several instances of loops, or adjectives with an edge that starts and ends at the same node. There are eight loops spanning seven adjective nodes: ‘tight’, ‘sore’, ‘cold’, ‘intense’, ‘hurting’, ‘tingling’, ‘itching’. These adjectives were each found in a context like the ones given in (25) and (26).

(25) **itchy but not itchy** to where I am scratching-type feeling
(26) my muscles are still **tight but not** nearly as **tight** as they used to be

These cases demonstrate that adjectives can not only be part of scales, but themselves have scalable properties. Though the statement ‘itchy but not itchy’ is (at least, on its face) self-contradictory, the sentence given in (25) compares ‘itchy’ to ‘itchy to where I am scratching-type feeling,’ which appears to convey a stronger sensation than just ‘itchy.’ Further research could attempt to account for the effect of modifiers and predicates on adjective intensity, particularly in contexts where an adjective is compared to a different intensity version of itself.

***Disconnected Nodes.*** Lastly, another starkly apparent visual on the graph is the separation between the ‘tight’ and ‘squeezing’ nodes and the rest of the adjectives. While all other adjectives are connected to more than one other adjective node, whether directly or via paths through other nodes, ‘tight’ points only to ‘squeezing’ (and itself), and ‘squeezing’ has no additional emanating arrows. In terms of frequency, both ‘tight’ and ‘squeezing’ are in the bottom third of weak-strong pair adjectives, though ‘tight’ ranks higher up in overall corpus adjective frequency. Again, considering the frequency of certain adjectives in the general lexicon would be an interesting follow-up to this research, and could perhaps address phenomenon such as the isolated [tight, squeezing] pair.

## 7. RESULTS AND DISCUSSION

As demonstrated by the analysis presented in Section 6, there is not a neat, linear relationship between all the adjectives collected. This, however, does not automatically support the null hypothesis $H0_a$, that people do not use adjectives in a predictable and scalable manner. Natural language is dynamic and fluid. So, while the overall picture is somewhat messy, the more informative process is attempting to find patterns within the complexity. Specifically, the graph is considered for its agreement with the intensity scales presented by the MPQ. From the final scales constructed in this research (found in Figure 1 in Section 5.3), there are 17 adjective relationships that can be defined by comparing each element in a given scale to all the other elements in that scale. For example, in the scale <annoying, intense, unbearable>, it is defined not only that *unbearable* is stronger than *intense,* but also that *unbearable* is stronger than *annoying*, etc. These relationships could also be calculated for the same adjectives in the MPQ. Of these 17 adjective relationships, the scales developed in this research demonstrated 58.8% agreement with the MPQ. That is, the HealingWell scales correctly predicted the relative strength of two adjectives as defined by the MPQ 10 out of 17 times. Of the remaining seven, four were incorrectly predicted. For example, where the MPQ defines *hot* as less intense than *burning*, the HealingWell scales defined *hot* as more intense than *burning*. The remaining three were inconclusive, with the HealingWell scales unable to predict the relative intensity between two adjectives. This was the case for the scale <dull/sore/aching, heavy, hurting>, where *dull* and *sore* were both defined as less intense than *heavy*, but no information was obtainable for the intensity of *dull* relative to *sore* and vice versa.

Since the expected agreement due to chance for these two datasets is 50% (for each two adjectives compared, they were either both in the same order as the MPQ or both in the opposite order), the HealingWell scales demonstrate an above chance level agreement with the MPQ, though only by 8.8%. I conducted a two-sample t-test to determine the statistical significance of the difference between chance and observed agreement with the MPQ orderings. The t-statistic was not significant at the .05 alpha level, with t(142)=.682, p=.497. However, considering the small sample size, a larger confidence level might be appropriate, and the t-statistic does show significance at the .5 alpha level. At the .5 alpha level, the null hypothesis that people do not use adjectives predictably can be rejected. This suggests that there is support for hypothesis H1, that there is some predictability to the way in which people use adjectives to describe their pain.

Even with support for hypothesis H1, it appears that the pattern of scalable adjective use is more nuanced than the MPQ defines it, with the intensity of adjectives not always entirely pinpointable. Furthermore, the HealingWell scales were produced using the categories defined by the MPQ. As discussed above, there is reason to believe that the MPQ categories are not ideally divided based on how people use the MPQ adjectives. Since Horn scales are only reasonable for adjectives that differ in intensity rather than semantic category, further analysis would be needed to consider which adjectival scales should be constructed from the weak-strong pairs collected.
Finally, this research also provided a preliminary investigation into Hypothesis 2, on potential differences in adjective usage between different types of chronic pain sufferers. At the start of this research, I hypothesized that there might be a difference in adjective frequency across different subcategories of chronic pain. In section 4.5, I compared frequency data across different corpora, and between the different topics in the created HealingWell corpus. While RA was an outlier for some adjective frequencies, the differences are not appreciably significant, due to the small sample size for the lower-frequency adjectives (less than 10 or so occurrences in a multi-million-word corpus). Given the overall similarity of adjective frequencies across corpora and forum topics, this research provides preliminary support for $H0_b$, that different categories of chronic pain are not associated with specific adjective pain descriptors.

## 8. Conclusion

This study was conducted to test the concepts behind the design of the McGill Pain Questionnaire, a clinical tool for assessing pain quality. Whereas previous work had recreated the MPQ by eliciting survey data, this research attempted to reconstruct the adjective intensity relationships defined by the MPQ by looking only at sentence construction patterns in spontaneous speech. Spontaneous speech more closely approximates the ways in which people understand and use adjectives as pain descriptors. In considering the field of pain assessment, the goal is to facilitate communication of pain from patient to physician. As such, to judge the efficacy of an adjective-based pain questionnaire, it is important to understand how patients describe their pain without prescribed frameworks like the MPQ. The downside to using a natural corpus, like the HealingWell corpus developed in this research, compared to survey data is the unpredictability of the data. Here, very specific patterns of adjective use were required, which were only found in small quantities in the forum corpus. The small sample sizes limit the power of the conclusions drawn in this research. Still, the successful reconstruction of adjective intensity scales from partial orderings, however limited, will hopefully begin an insightful conversation on the structure of current clinical pain questionnaires. For the adjectival pain questionnaire to be a valuable clinical tool, the patient's understanding and use of each adjective must align with the meaning prescribed and interpreted by the receiving physician; in other words, the patient's pain must be communicable through adjectives. This research suggests that, given the right division of adjectives by category, it is possible to predict the relative intensities of adjectives within a given category to some degree. Further research will be needed to determine which categorizations of adjectives are best, and how to find them.

## ETHICAL CONSIDERATIONS

All the collected data comes from HealingWell.com, whose privacy policy states that all forums are public and accessible to guests and search engines [13]. No identifying information—including name, age, gender, or location—was collected in connection to any of the forum data, to protect the privacy and identities of the post authors.

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Fellbaum, for her mentorship and guidance throughout the research and writing process.

I also owe a thank you to my boyfriend Ariel, for his patience and assistance in understanding and optimizing the computer science concepts used for this research.

Finally, I would like to thank my mom for her thoughtful editing and helpful brainstorming.

## REFERENCES

[1] Melzack, R. (1975). "The McGill Pain Questionnaire: Major Properties and Scoring Methods" in *Pain*, vol. 1, no. 3, pp. 277–299. https://doi.org/10.1016/0304-3959(75)90044-5

[2] Melzack, R. & Raja, S.N., (2005), "The McGill Pain Questionnaire: From Description to Measurement," *Anaesthesiology*, vol 103, no. 1, pp. 199-202. https://doi.org/10.1097/00000542-200507000-00028

[3] Frederiksen, L.W., Lynd, R.S., & Ross, J. (1978), "Methodology in the Measurement of Pain," *Behavior Therapy*, vol. 9, no. 3, pp. 486-488. https://doi.org/10.1016/S0005-7894(78)80095-1

[4] Haefeli, M., & Elfering, A., (2006), "Pain assessment," European Spine Journal, vol. 15, pp.17-24. https://doi.org/10.1007/s00586-005-1044-x

[5] Reading, A. E., Everitt, B. S., & Sledmere, C. M. (1982), "The McGill Pain Questionnaire: A replication of its construction," *British Journal of Clinical Psychology*, vol. 21, no. 4, pp. 339–349. https://doi.org/10.1111/j.2044-8260.1982.tb00571.x

[6] Kremer, E. & Atkinson, H.J. (1981), "Pain Measurement: Construct Validity of the Affective Dimension of the McGill Pain Questionnaire with Chronis Benign Pain Patients" in *Pain*, vol. 11, no. 1, pp. 93-100.

[7] Kenneth A. Holroyd et al. (1992), "A Multi-Center Evaluation of the McGill Pain Questionnaire: Results from More than 1700 Chronic Pain Patients," *Pain*, vol. 48, no. 3, pp. 301-311.

[8] Kahl, C. & Cleland, J.A. (2005), "Visual Analog Scale, Numeric Pain Rating Scale, and the McGill Pain Questionnaire: An Overview of Psychometric Properties," *Physical Therapy Reviews*, vol. 10, no. 2, pp. 123-128.

[9] Farkas, D. LIN311, "Varieties of Meaning: Semantic and Pragmatic Approaches," Class Lecture, Meeting 1. Program in Linguistics, Princeton University, Princeton, NJ, Feb. 4, 2020.

[10] Horn, L., (1972), "On the Semantic Properties of Logical Operators in English". Ph.D. dissertation, Philosophy in Linguistics, University of California, Los Angeles.

[11] Grice, P., (1975), "Logic and Conversation" in *Speech acts*, 5th ed., Academic Press; University College London, pp. 41–58.

[12] Farkas, D. LIN311, "Varieties of Meaning: Semantic and Pragmatic Approaches," Class Lecture, Meeting 2. Program in Linguistics, Princeton University, Princeton, NJ, Feb. 6, 2020

[13] HealingWell, "Chronic Illness Support," *healingwell.com*, 1997-2021. [Online]. Available: https://www.healingwell.com/community/default.aspx?f=16. [Accessed Mar. 13, 2021].

[14] Mayo Clinic Staff, (2020, October 07). "Fibromyalgia: Symptoms & Causes," *mayoclinic.org*. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/fibromyalgia/symptoms-causes/syc-20354780. [Accessed June 12, 2021].

[15] Mayo Clinic Staff, (2019, March 07). "Rheumatoid Arthritis: Symptoms & Causes," *mayoclinic.org*. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/rheumatoid-arthritis/symptoms-causes/syc-20353648. [Accessed June 12, 2021].

[16] Richardson, Leonard, (2021), BeautifulSoup Documentation. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed Dec. 17, 2020].

[17] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc. [Online]. Available: http://www.nltk.org/book/. [Accessed Dec. 10, 2020].

[18] Barrus, T. (2021), pyspellchecker (Version 0.6.2), [Open-source package]. Available: https://pypi.org/project/pyspellchecker/. [Accessed Feb. 11, 2021].

[19] Francis, W.N., & Kucera, H., (1979). "Brown Corpus Manual," Department of Linguistics, Brown University, Providence, Rhode Island. [Online]. Available: http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM. [Accessed June. 15, 2021].

[20] Davies, Mark. (2008) "The Corpus of Contemporary American English," *english-corpora.org*. [Online]. Available: https://www.english-corpora.org/coca/. [Accessed June. 15, 2021].

[21] Han, J., Kambler, M., & Pei, J. (2012). "Getting to Know Your Data" in *Data Mining*, Elsevier, Inc, pp. 39-82. https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[22] Sheinman, V., Fellbaum, C., Julien, I., Schulam, P. & Tokunaga, T., (2013), "Large, Huge, or Gigantic? Identifying and Encoding Intensity Relations Among Adjectives in WordNet," *Language Resources and Evaluation*, vol. 47, no. 3, pp. 797-816.

[23] Aric A. Hagberg, Daniel A. Schult & Pieter J. Swart, (Aug. 2008), "Exploring Network Structure, Dynamics, and Function Using NetworkX", in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), Pasadena, CA, pp. 11–15.

[24] Bank, Sebastian (2013-2021), "graphviz Documentation," [Online]. Available: https://graphviz.readthedocs.io/en/stable/manual.html. [Accessed Apr. 23, 2021].

[25] Princeton University (2010), "WordNet Search." *WordNet*. Princeton University.

## AUTHOR

**Miriam Stern** is an undergraduate student at Princeton University entering her fourth and final year in the Program in Linguistics. She is also completing the pre-med track and will be attending Sydney Kimmel Medical College at Thomas Jefferson University upon completion of her undergraduate studies. Miriam hopes to continue traversing the intersection between language and medicine in her future endeavours.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# A SELF-SUPERVISED TIBETAN-CHINESE VOCABULARY ALIGNMENT METHOD BASED ON ADVERSARIAL LEARNING

Enshuai Hou and Jie zhu

Tibet University, Lhasa, Tibet, China

## ABSTRACT

*Tibetan is a low-resource language. In order to alleviate the shortage of parallel corpus between Tibetan and Chinese, this paper uses two monolingual corpora and a small number of seed dictionaries to learn the semi-supervised method with seed dictionaries and self-supervised adversarial training method through the similarity calculation of word clusters in different embedded spaces and puts forward an improved self-supervised adversarial learning method of Tibetan and Chinese monolingual data alignment only. The experimental results are as follows. First, the experimental results of Tibetan syllables Chinese characters are not good, which reflects the weak semantic correlation between Tibetan syllables and Chinese characters; second, the seed dictionary of semi-supervised method made before 10 predicted word accuracy of 66.5 (Tibetan - Chinese) and 74.8 (Chinese - Tibetan) results, to improve the self-supervision methods in both language directions have reached 53.5 accuracy.*

## KEYWORDS

*Tibetan; Word alignment; Without supervision; adversarial training.*

## 1. INTRODUCTION

Tibetan is a low-resource minority language, and the available Tibetan-Chinese sentences pair corpus is relatively scarce compared to English, Chinese, etc. However, research on tasks such as Tibetan-Chinese bilingual machine translation [1,2] requires a large number of bilingual comparisons. Compare Tibetan and Chinese bilingual corpus, large monolingual data more readily available. Extracting similar words which have semantic information from the Chinese and Tibetan monolingual corpus generates a comparison dictionary. This work can alleviate the need for bilingual comparison data in tasks such as machine translation. Based on Harris' [3] distribution hypothesis, words with similar contexts have similar semantics, and word vectors can reflect this distribution relationship to a large extent. The word vectors of similar words are relatively nearby to the embedding space and come from word clusters of different languages. Word clusters from different languages have similar distributions in different embedding spaces [4,5]. In this paper, semi-supervised, self-supervision and improved the self-supervision threekindsofTibetan-Chinese bilingual glossary of methods, and the use of syllables and words two kinds of segmentation granularity experiment. In semi-supervised alignment method of mapping we use a seed dictionary to map,then spread throughout the semantic space; self-supervision alignment method, against networks [6] to learn mapping, by the mapping the representative word embedded mapping space to for aligned Tibetan Chinese vocabulary: The improved Tibetan-Chinese self- supervised vocabulary first uses a self-supervised method to learn to map, and then uses part of the high-frequency word pairs generated by this mapping as the seed dictionary, and iteratively improves the semi-supervised method to obtain the final

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 277-285, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111422

Tibetan-Chinese aligned dictionary. The research of Chinese and Tibetan self-supervision vocabulary alignment method can effectively alleviate the need for bilingual data onto research and it has a positive meaning involving Chinese and Tibetan.

## 2. RELATED WORKS

There are many relevant pieces of research on obtaining cross-language vocabulary pairs from monolingual data onto different languages without using a grand number of parallel corpora. In 2013, Mikolov et al. [7] first used the hypothesis that words with different languages in similar contexts are similar to learn a linear mapping from the source of the target embedding space, using 5000 pairs of parallel words as anchor points for training, and evaluate their methods in a word translation task. The research of Xing et al. [8] in 2015 showed that compared with the method of Mikolov et al, the result can be improved by implementing orthogonal constraints on the linear mapping. In 2017, Smith et al. [9] used the same characters to construct a bilingual dictionary in an attempt to reduce the dependence on supervised bilingual dictionaries. In 2017, Artetxe et al. [10] initialized bilingual dictionaries with aligned numbers, and then gradually aligned the embedding space using an iterative method. However, their model can only be applied to languages with similar alphabetic languages due to the shared alphabet. Without using parallel corpus, Zhang et al. [11] began to try to use adversarial methods to obtain cross-lingual word embedding. Alexis et al. [12] used the adversarial training method to learn a linear mapping from the source language to the target language space, then used the Procrustes method of fine-tuning, and proposed a bilingual vocabulary similarity measurement method of cross-domain similarity local scaling. In Tibetan and Chinese language pairs, there is also the problem of a lack of data resources. Methods such as back translation [13] can expand the data and construct pseudo-parallel corpus, but the more demand for parallel sentence pairs and problems of quality cannot be avoided.

## 3. SELF-SUPERVISED TIBETAN-CHINESE VOCABULARY ALIGNMENT METHOD

This chapter mainly describes the Tibetan-Chinese vocabulary alignment method of the Tibetan-Chinese language differences, the Tibetan-Chinese vocabulary similarity measurement method, the semi-supervised method with a seed dictionary, the self-supervised method, and the improved self-supervised method. According to the research content of this article, firstly, it introduces the differences between Tibetan and Chinese languages from the question of whether the constituent elements of Tibetan and Chinese languages require segmenting. Secondly, it introduces the method of measuring similarity between Tibetan and Chinese vocabularies and their use. Then, from the perspective of the model, the semi-supervised seed dictionary method, the self-supervised Tibetan-Chinese vocabulary alignment method based on adversarial training, and the improved method based on self-supervised alignment with the semi-supervised method is described.

### 3.1. Differences between Tibetan and Chinese

Tibetan is a phonetic script and has its own special language organization. The smallest morpheme in Tibetan is Tibetan syllable, and the smallest semantic unit is Tibetan words. A Tibetan syllable is composed of one or more Tibetan characters, and Tibetan words can be composed of one or more syllables, but there is no obvious division of words, and there is still a phenomenon of deflation [14]. The representation of the same thing in different dialects is also different, causing many processing difficulties and increasing the difficulty of labeling Tibetan data.

Chinese is a square text, the semantic unit is a word, one or more Chinese characters can form a Chinese vocabulary, there are no separators between words, and there are many polysemous phenomena in a word. Comparing Tibetan and Chinese, two scripts with the same semantic length, in a computer, Tibetan requires more storage space, etc.

## 3.2. Similarity Measurement Method of Tibetan and Chinese Vocabulary

In order to measure the similarity between Tibetan and Chinese words, this paper uses Cross-domain Similarity Local Scaling(CSLS) as a measure of similarity between Tibetan and Chinese words. This method is based on the improvement of the K-nearest neighbor method, and solves the problem of the K-nearest neighbor method of high-dimensional space: the nearest neighbor is asymmetric. In the high-dimensional space, some words are the nearest neighbors of many words, and some words are not neighbors ofany words. The nearest neighbor and the average similarity of the nearest neighboris introduced as a penalty factor of the CSLS method, which improves the local accuracy.

The CSLS method is defined as follows: Suppose there is a two-part neighborhood graph in the vector space of two languages, where each word is connected to $K$ nearest neighbors in the other language. Then respectively calculate the cosine similarity between the word and the $K$ nearest neighbors, and finally take the average value to obtain the average similarity $r$ between a vocabulary and adjacent vocabulary in another language.

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K}\sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t) \,(1)$$

Combining the two conversion directions of the source language to the target language and the target language to the source language, the cosine similarity is combined with r in the two directions as the penalty factor to form the similarity measurement method CSLS:

$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t)(2)$$

This paper will use this similarity measurement method in two aspects: one is to use it in the process of generating bilingual alignment dictionaries; the other is to use it in the result evaluation index, taking the accuracy of the first N similar words, that is, P@N as the evaluation index.

## 3.3. Semi-supervised Method with Seed Dictionary

For Tibetan and Chinese, use the monolingual corpus to train word vectors, and obtain two sets of vectors with dictionary size n, m. Source language vector is defined as: $\{x_1, \dots, x_n\}$, and Target language vector is defined as: $\{y_1, \dots, y_m\}$.

Based on these two sets of vectors, a cross-language vocabulary alignment method is learned. First, use the word vectors generated by the two monolingual corpora of Tibetan and Chinese, and use some word pairs $\{x_i, y_i\}_{i \in \{1,n\}}$, $x \in X, y \in Y$ in the seed dictionary, as n pairs anchor point, by minimizing $||WX - Y||$, learns the mapping $W$, as shown in formula 3.

$$W^* = \underset{\mathrm{W} \in M_d(\mathbb{R})}{argmin} \, ||WX - Y|| \; (3)$$

Where $d$ is the dimension of the word vector, and $W$ belongs to a $d \times d$ real matrix $M_d(\mathbb{R})$Xand $Y$ are $d \times n$dictionary embedding matrices to be aligned. Secondly, according to the research of Xing et al. [8], orthogonal constraints are implemented on $W$ to improve the results, and $W$ is an orthogonal matrix through singular value decomposition constraints, as shown in formula 4. Among it d is the dimension of the word vector, and $W$ belongs to a matrix of real numbers. And are to be aligned with the size of the word typical embedded matrix.

$$W^* = \underset{W \in O_d(\mathbb{R})}{argmin} \|WX - Y\|_F = UV^T, U\sum V^T = SVD(XY^T) \quad (4)$$

Through iterative training in two alignment directions, update $W$, minimize the difference between $WX$ and $Y$, and use mapping $W$ to align vocabulary to generate Tibetan-Chinese bilingual word pairs $\{Wx_i, y_i\}_{i\in\{1,t\}(t<m,t<n)}$, changes the translation direction and repeat the experiment, use the following method to generate two translation direction dictionaries.

In the dictionary generation process, on $\{Wx_1, \dots, Wx_n\}$and $\{y_1, \dots, y_m\}$, uses the CSLS method to calculate the CSLS value of K neighbors in two directions. Based on the seed dictionary, the distance between the two languages dictionary is updated for the closest word pair, and a new aligned dictionary is finally generated.

### 3.4. Self-supervision Methods

The self-supervised method uses a generative adversarial network, defines the discriminator as a fully connected network, and the generator is a randomly initialized linear mapping $W$, and learns the latent space between Tibetan and Chinese through the training of the discriminator and generator. The discriminator is used to distinguish the elements from $WX = \{Wx_1, \dots, Wx_n\}$and $Y$. The role of the discriminator is to distinguish the two embedding sources. The mapping function $W$ to be learned by the generator makes it difficult for the discriminator to distinguish whether the word embedding is $Wx \ (x \in X)$ or $y \ (y \in Y)$ . .

The parameter that defines the discriminator network is $\theta_D$. When we assume z is a word embedding vector of unknown source, $P_{\theta_D}(source = 1|z)$ represents the probability that the discriminator considers the vector z to be the source language embedding$P_{\theta_D}(source = 0|z)$means that the discriminator considers the probability that the vector z is embedded in the target language. Here, the cross-entropy loss is used as the loss of the discriminator. The formula of loss function is as follows:

$$L_D(\theta_D \mid W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{ source } = 1 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{ source } = 0 \mid y_i) \quad (5)$$

Correspondingly in self-supervised training, for generator mapping $W$, its loss is defined as follows:

$$L_W(W \mid \theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{source } = 0 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(source = 1 \mid y_i) \quad (6)$$

This article conducts the training of the adversarial network according to the standard adversarial network training process described by Goodfellow et al. [6] For a given input sample, the discriminator and the mapping matrix $W$ are updated using the stochastic gradient descent method to minimize $L_D$ and $L_W$ and finally make the two loss functions no longer drop. At the same time, the orthogonal constraint is added when updating $W$ during the training process, and it is used alternately with gradient descent. The constraint formula is shown in formula 7. Orthogonal constraints can maintain the dot product of the vector, ensure the quality of the word embedding corresponding to the language, and make the training process more robust. The value of $\beta$ is generally 0.001 to ensure that $W$ can almost always remain orthogonal.

$$W \leftarrow (1+\beta)W - \beta(WW^T)W \ (7)$$

Finally, a self-supervised Tibetan-Chinese bilingual aligned dictionary is generated using the dictionary generation method in the semi-supervised method with seed dictionary mentioned before.

### 3.5. Improve Self-supervised Adversarial Learning Method

In the self-supervised adversarial method, the mapping relationship of high-frequency words can better reflect the global mapping relationship. From the self-supervised adversarial method generation dictionary, select $s$ words pairs with a higher frequency of occurrence 〖 $\{Wx_i, y_i\}_{i\in\{1,s\}}$ as anchor points, and then use the semi-supervised method to improve Training, iteratively updates the mapping $W$, extend the mapping to the lower frequency vocabulary domain, and generate a new alignment dictionary. The dictionary generation process is the same as the semi-supervised method of the seed dictionary.

## 4. EXPERIMENT AND ANALYSIS

### 4.1. Data Collection

Two separate corpora of Tibetan and Chinese were collected from the Internet, each with more than 35,000 sentences. The two-grained segmentation processing was performed on Tibetan and Chinese respectively, and experiments on the syllable size and word size of both Tibetan and Chinese characters were carried out.

The two dictionary pairs, Tibetan-Chinese and Chinese-Tibetan, are constructed artificially as the anchor point and test dictionary of the semi-supervised training method. The word granularity dictionary size is 10000, and the syllable granularity dictionary size is 3000.

### 4.2. Parameters Settings

The word vector training adopts the skip-gram model of the fast text model, and the vector dimension is 300. The training framework uses Pytorch, and the specific parameter settings are shown in Table 1 and Table 2. The semi-supervised method and the improved part of the improved self-supervised method use consistent parameters.

Table 1. Semi-supervised and improved self-supervised parameter Settings

| **parameter** | value |
|---|---|
| **Seed dictionary size (syllables)** | 1500 |
| **Seed dictionary size (words)** | 5000 |
| **Test set size (syllables)** | 500 |
| **Test set size (words)** | 1500 |
| **Number of iterations** | 5 |
| **Word vector normalization processing** | True |
| **Generate a dictionary word on the number of** | 50000 |

Table 2. Self-supervised adversarial training parameter setting

| **parameter** | value | **parameter** | value |
|---|---|---|---|
| **Test set size (syllables)** | 500 | Number of training cycles per training | 100000 |
| **Test set size (words)** | 1500 | Discriminator network layer number | 2 |
| **Word vector normalization processing** | True | Number of single-layer nodes | 2048 |
| **Generate dictionary word pairs** | 50000 | Activation function | LeakReLU |
| **Training times** | 5 | Network input discard rate | 0.1 |
| **Orthogonalization parameters** | 0.001 | Training batch size | 32 |

### 4.3. Results Analysis

This paper conducts two methods of experiments on Tibetan and Chinese monolingual corpora with syllable granularity and word granularity in two alignment directions. One is the accuracy of the word granularity experiment under (1, 5, 10) candidate words, which is P@N. Experiment, and compare the experimental effects of Tibetan → Chinese (Ti-Zh) and Chinese → Tibet (Zh-Ti); the second is the accuracy of the word granularity under (1, 5, 10) candidate words, P@N's experiment, and compared the experimental effects of Tibet→Chinese (Ti-Zh) and Chinese → Tibet (Zh-Ti).In tables, these use Semi-sup, Self-sup, Self-sup-re to Means semi-supervised method,self-supervised adversarial and the refine method of self-supervised adversarial.

In the syllable granularity experiment, as shown in Table 3, as the number of candidate words increases, the accuracy of the semi-supervised model P@N gradually increases, but the overall is badalthough the self-supervised method has an increasing trend, In two directions it is difficult to learn the semantics of embedding space in this direction; after the improvement, the effect cannot be improved, and even the performance has declined in several results, showing a kind of disorder, which proves the weak semantic connection between Tibetan syllables and Chinese characters.

Table 3. Syllable granularity the value of P@N in both alignment directions

| | Ti-Zh | | | Zh-Ti | | |
|---|---|---|---|---|---|---|
| | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 |
| **Semi-sup** | 11.8 | 24.8 | 30.9 | 14.4 | 24.2 | 28.6 |
| **Self-sup** | 0.2 | 1.0 | 1.0 | 0.2 | 0.4 | 1.0 |
| **Self-sup-re** | 0 | 1.0 | 1.6 | 0 | 0.2 | 0.6 |

In terms of word granularity experiments, as shown in Table 4, compared to the self-supervised method, the semi-supervised method has achieved good results, and the improved Tibetan-Chinese self-supervised alignment method has achieved good results in the direction of Tibetan →Chinese. Both reached 53.5. In this process, the improved training played a great role and significantly improved the experimental effect. Although the improved self-supervised method has achieved a weaker effect than the semi-supervised method, it is of positive significance because it does not use the contrast dictionary for training.

Table 4. P@N values in the two alignment directions under the granularity of Tibetan and Chinese words

| | Ti-Zh | | | Zh-Ti | | |
|---|---|---|---|---|---|---|
| | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 |
| **Semi-sup** | 48.5 | 62.7 | 66.5 | 55.7 | 69.8 | 74.8 |
| **Self-sup** | 12.7 | 23.7 | 29.2 | 8.4 | 16.8 | 21.7 |
| **Selfs-up-re** | 25.4 | 46.3 | 53.5 | 27.5 | 47.7 | 53.5 |

After training, the word vectors corresponding to the partially aligned vocabulary of the two languages of Tibetan and Chinese are subjected to principal component analysis (PCA), and the two-dimensional vectors are visualized on the plane. The result is shown in figure 1. It can be seen from the figure that Tibetan and Chinese words have similar meanings in similar positions in space.

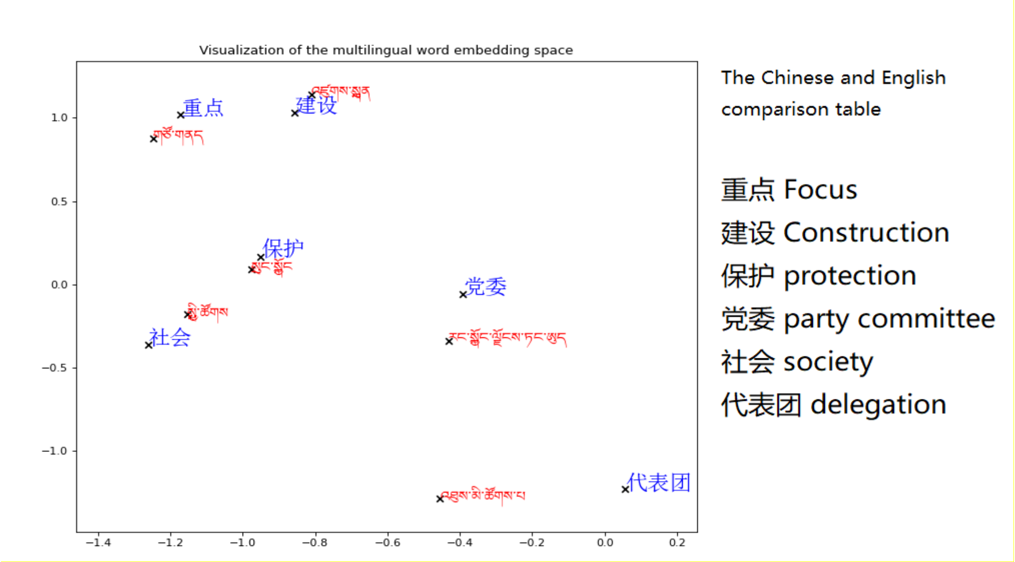


Fig 1. PCA visualization of partial results

## 5. CONCLUSIONS

In order to alleviate the need for parallel corpus for bilingual sentences in tasks such as Tibetan-Chinese translation, this paper uses the assumption that words in similar contexts are similar in different languages, using semi-supervised and self-supervised methods from Tibetan-Chinese monolingual data Extract aligned Tibetan-Chinese bilingual vocabulary and construct a bilingual aligned dictionary. First, the experiment of syllable granularity proves the weak correlation between Tibetan syllables and the semantics of Chinese characters. Second, improved self-supervision methods have achieved relatively effective results in the experiment of word granularity. The research in this article can alleviate the scarcity of Tibetan-Chinese parallel corpus data and provide a good start for unsupervised Tibetan-Chinese machine translation.

## ACKNOWLEDGEMENTS

Foundation Project: "Construction of Tibetan and Chinese Bidirectional Machine Translation Platform for the Inheritance and Development of Tibetan Language" and " Tibetan language training technology and Application".

## REFERENCES

[1] Artetxe M, Labaka G, Agirre E. Bilingual lexicon induction through unsupervised machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,Florence,Italy,2019:5002-5007.

[2] Renqing Dongzhu, Toudan Cairang, Nima Ttashi. A review of the study of Chinese-Tibetan machine translation[J].China Tibetology,2019(04):222-226.

[3] Zellig S Harris. Distributional structure. Word, 10(2-3):146–162, 1954.

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. Proceedings of EMNLP, 14:1532–1543, 2014.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, AaronCourville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, pp. 2672–2680, 2014.

[7] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168, 2013b.

[8] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. Proceedings of NAACL, 2015

[9] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. International Conference on Learning Representations, 2017.

[10] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 451–462. Association for Computational Linguistics, 2017.

[11] Meng Zhang, Y ang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2017b.

[12] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jegou. Word translation without parallel ´ data.arXiv:1710.04087, 2017.

[13] Cizhen Jiacuo, Sangjie Duanzhu, Maosong Sun, et al. Research on Tibetan-Chinese machine translation method based on iterative back translation strategy[J].Journal of Chinese Information Processing,2020,34(11):67-73+83.

[14] Caizhijie. Recognition of compact words in Tibetan word segmentation system [J].Journal of Chinese Information Processing, 2009, 23(01):35-37+43.

## AUTHORS

**Enshuai Hou** 1996.09 Postgraduate in Tibet University Lhasa Tibet China Direction: Tibetan-chinese machine translation.

**Jie Zhu** 1973.11Professor doctoral supervisorin Tibet University Lhasa Tibet China Direction: Natural language processing (NLP), Tibetan information processing, Data Mining and artificial intelligence.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# INDUSTRIAL BIG DATA ANALYTICS AND CYBER-PHYSICAL SYSTEMS FOR FUTURE MAINTENANCE & SERVICE INNOVATION

Temitope O Awodiji

Computer Information Science Personnel,
California Miramar University, California, USA.

## *ABSTRACT*

*Based on Information and Communication Technologies (ICT) fast advancement and the integration of advanced analytics into manufacturing, products, and services, several industries face new opportunities and at the identical time challenges of maintaining their ability and market desires. Such integration, that is termed Cyber-physical Systems (CPS), is remodeling the industry into a future level. CPS facilitates the systematic conversion of big data into information that reveals invisible patterns of deterioration and inefficiency and leads to better decision-making. This project focuses on existing trends within the development of industrial huge information analytics and CPS. Then it, in brief, discusses a system architecture for applying CPS in manufacturing referred to as 5C. The 5C architecture, comprises necessary steps to totally integrate cyber-physical systems within the manufacturing industry.*

## *KEYWORDS*

*Information and Communication Technologies (ICT), Big Data, Analytic, Data, Data Science, Data Architecture, Cyber Physical Systems, Integration.*

## 1. INTRODUCTION

Unmet needs in today's business During the last decade's manufacturers and service providers have taken an important step to improve the standard of products and services and to optimize their processes to survive in the competition and to react to market demands. Customer orientation, value creation, and service orientation quality oriented This development have led to the development of Prognostics and Management (PHM). Information and knowledge about the invisible patterns of asset depreciation. and the inconsistencies and inefficiencies of the processes. These patterns are invisible until an error occurs [(Lee, Lapira, Bagheri, Kao, (2013)] posits that the invention of such underlying patterns avoids the expensive failures and unplanned downtime of machinery. Such a maintenance scheme results in larger quality sustainability near-zero breakdown.

## 2. CPS 5C LEVEL ARCHITECTURE

In recent years, the rapid advancement of information and communication technologies (ICT) has accelerated the use of advanced sensors, data collection devices, wireless communication devices, and remote computing solutions. The integration of advanced analytics with communication technologies in close connection with physical machines is known as Cyber-Physical Systems (CPS). According to Shi, Wan, Yan, and Suo (2011), “An Overview of Cyber-

DOI: 10.5121/csit.2021.111423

Physical Systems”. Since the inception of its concept, CPS has been an ever-growing terminology in today's evolving industry. Predicting Potential Failures By implementing this predictive analysis in addition to a support system of your choice, the right services can be requested, and action can be taken to maximize the uptime, productivity, and efficiency of industrial systems. As the central hub for data management at the fleet level, CPS plays a crucial role in achieving the goals mentioned above.

## 3. Internet of Things and Evolution of Industries

According to The Internet of Things (IoT), Kopetz [(2011)] posits that the internet of things. In real-time systems will gather, sort, synchronize and organize the data from totally different sources within a manufacturing plant or business. It provides a seamless, connected data management platform with real-time streaming and processing capabilities. This platform offers the possibility to implement a predictive analysis of big data for the transformation of data into information, knowledge, and actions through a CPS structure. Data to Action has the potential to add value to different parts of a business chain. For example, valuable data on hidden deterioration or inefficiency patterns between machines or production processes can lead to intelligent and effective maintenance options that avoid costly downtime and unplanned downtime.

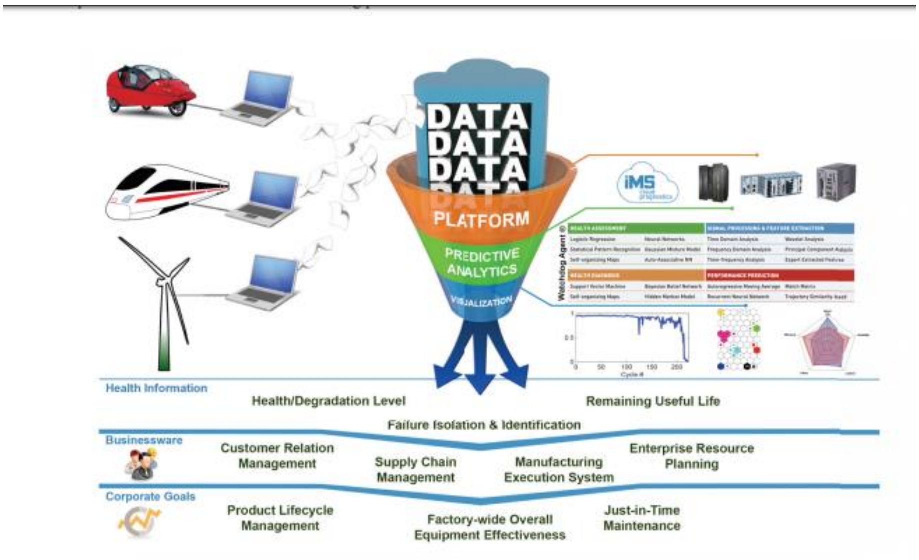


Figure 1. The framework of big data analytics platform for predictive manufacturing [2]

Figure 1 provides a schematic view of how massive data analytics will produce value among completely different sections of industries. Within the next section, the structure of CPS alongside its implementation aspects is mentioned.

## 4. Cyber-Physical Systems

The CPS structure, projected in Lee, J., B., Bagheri, H. A., Kao (2013), consists of five levels specifically 5C architecture. This framework provides a guide for the development of CPS for industrial applications. This CPS structure consists of 2 main components:

1) the advanced connectivity that ensures real-time information streamlining from the physical space to cyberspace and feedback from the cyberspace; and

2) Intelligent data analytics that constructs the cyberspace. The Projected 5C framework provides advances that show how a CPS system can be built from data acquisition to value creation. The framework of CPS at completely different levels is shown in Figure 2. The 5C structure consists of good affiliation, Data-to-info Conversion, Cyber, cognition and Configuration levels.

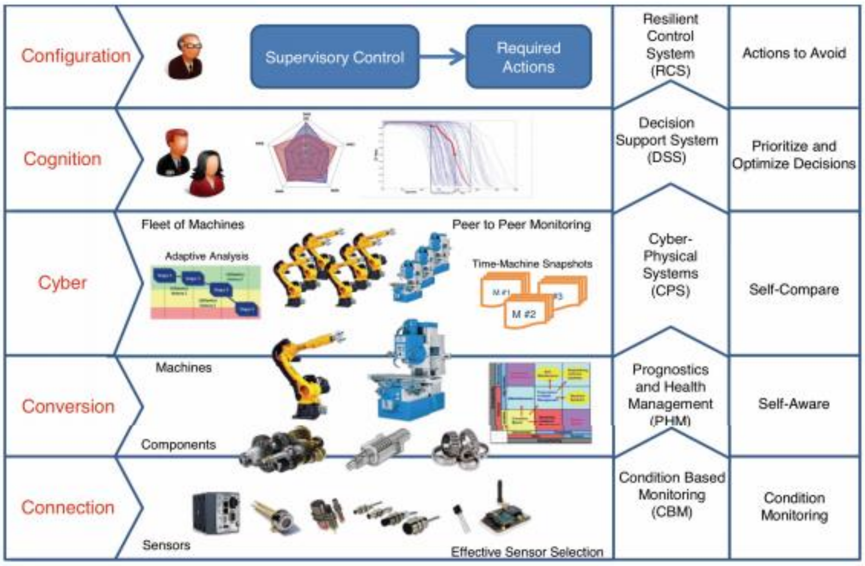


Figure 2. Framework of a cyber-physical system in different levels from data to action [5]

## 4.1. Smart Connection

This level consists of seamless and seamless strategies for managing data acquisition systems, optimizing data, and transferring it to the central server, selecting the right sensors, data sources and transfer protocols such as MT Connect. [Vijaya Raghavan, Sobel, Fox, Dornfeld and Wendorff (2008)]during this stage will have a great impact on the performance of the CPS within the following stages and on the quality and precision of the knowledge obtained by the system.

## 4.2. Data-To-Information Conversion

The core of such architecture is where the information is analyzed and remodeled into valuable data. Recently, there has been an in-depth focus on developing intelligent algorithms and data processing techniques. Such algorithms may be applied to numerous data sources, from machinery and process information to business and enterprise management data.

## 4.3. Cyber

The cyber level acts as the central information center in this architecture. Information is being pushed to that from each source and compiled to ascertain cyberspace. Having huge data gathered, specific analytics should be used to extract additional data that gives better insight into the status of individual machines among the fleet.

## 4.4. Cognition Implementing

Cognition Implementing CPS upon this level generates a thorough information of the monitored system. correct presentation of the acquired knowledge to professional users supports the right decision to be taken.

## 4.5. Configuration

The configuration level is the feedback from cyberspace to the physical area and acts as a supervisory controller to create self-configuring and self-adapting machines. This stage acts as a resilience control system (RCS) to use the corrective and preventive choices, that has been created within the knowledge level, to the monitored system.

Cyber-physical System-based smart Machine Machining processes within the manufacturing business represent a highly dynamic and complex scenario for condition-based maintenance (CBM) and PHM. A CNC machine will sometimes handle a good vary of materials with completely different hardness and geometric shapes and consequently needs different combos of machine and cutting parameters to operate. traditional PHM methods are sometimes developed for a restricted range of machine varieties operating and dealing conditions and so cannot be used to effectively handle a complete manufacturing floor where machines may be utilized under a good range of working regimes that cannot be sculptured comprehensively beforehand. As a result, a CPS framework with the projected 5C structure is developed for sawing processes. The developed Hz for machine tools may be used to process and analyses machining information, evaluate the health condition of important parts (e.g., tool cutter) and further improves the instrumentation efficiency and dependability by predicting forthcoming failures, scheduling maintenance beforehand and adaptive management.

In factories, manufacturing processes begin with sawing massive items of material into selected sizes. because of the upstream nature of the sawing method, the quality and speed of sowing influence the complete production and any error may be propagated to the subsequent steps and end in dangerous quality product. As shown in figure 3, within the connectivity level, data is acquired from machines through each add-on sensors and controller signals. additionally, to the add-on vibration, acoustic emission, temperature, and current sensors, twenty control variables like blade speed, cutting time and blade height are force out of the PLC controller to provide a transparent understanding of the working status of every machine. the information is currently processed within the industrial computer connected to every machine.

At the conversion level, the industrial computer also performs the feature extraction and the data preparation. The feature extraction consists of extracting standard time-domain and frequency domain options such as RMS, kurtosis, waveband energy percentage, etc. from vibration and acoustic signals The calculated functions are sent together with the machine status data via the Wi-Fi network to the cloud server, where the function values are managed and maintained within the database. At the cyber level, the cloud server uses an adaptive clustering technique.[Yang, Bagheri, Kao and Lee (2015)]posited to segment the blade performance history (from when a brand-new blade was installed to now) into discrete operating regimes supported the relative change of the options comparison to the conventional baseline and therefore the local noise distribution (figure 4). The adaptive clump technique compares the present values of the features with the baseline and historical operating regimes. It identifies the most appropriate cluster from the history to match with this operating condition. Although, if none is found, the algorithmic rule generates a replacement cluster as a brand-new operating regime and generate connected health models for that regime. Further, if a similar operating condition happens, the algorithmic rule has its signatures in memory and can mechanically cluster the new data into that specific operating regime [Yang, S., B., Bagheri, H. A., Kao, J., Lee (2015)].

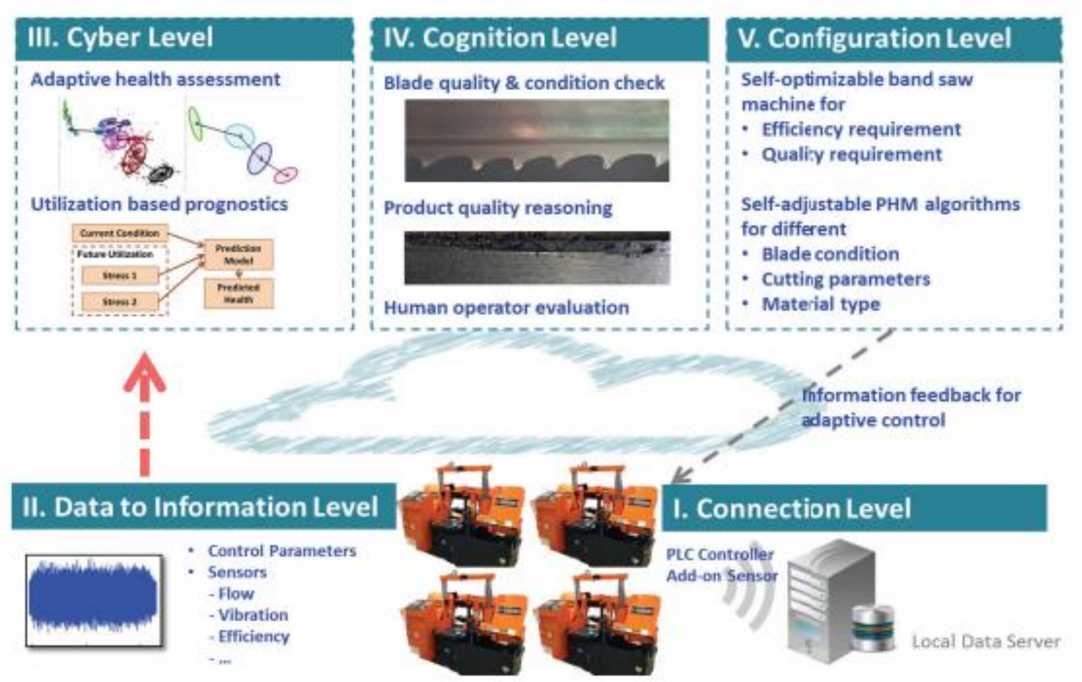


Figure 3. Overall Cyber Physical System setup for band saw Machines

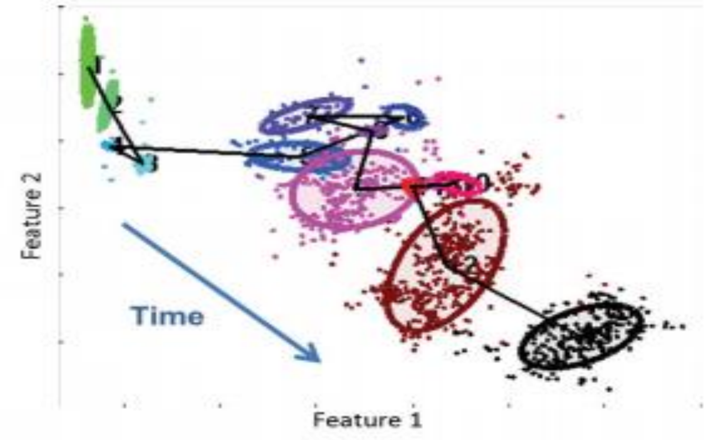


Figure 4. Adaptive health state recognition for blade condition (each color represent a cluster of feature values for a specific working regime)

Figure 5. User interface for real-time access of machine health information

The health stages are further utilized in the cognition level and configuration level for improvement functions. for instance, once the blade is new, a better cutting speed is used for top productivity without hampering the standard, while after certain quantity of degradation has been detected, a a lot of moderate cutting ought to be applied to make sure quality. To help such decision-making process, web and iOS-based user interfaces are additionally developed so the health data of every connected machine tool is accessed in real time (figure 5).

## 5. CONCLUSION

This paper explained current trends toward implementing cyber-physical systems within the manufacturing industry. As managing industrial massive data has become a difficult task for factories, coming up with a generic architecture for implementing CPS in manufacturing is important. The 5C architecture that is mentioned in this article will change and centralize data processing, health assessment, and prognostics. This architecture covers all necessary steps, from data acquisition and processing of information to presentation for the user and decision-making support.

However, the health data generated by the system are used for higher-level functions such as maintenance planning and optimized control for higher productivity and overall system reliability. The capabilities of the 5C architecture have shortly demonstrated through a case study of CNC saw machines. The case study shows the integration of the 5C architecture for process and managing a fleet of CNC sawing machines that are normally utilized in manufacturing. the present integration of the 5C CPS architecture is in its early stage, therefore, advancement in all five levels of the architecture is practical. The cyber level contains high development potential through the development of new algorithms for the fleet-related analysis of machine performance in distributed data management systems.

## REFERENCES

[1] Chan, F. T. S., H. C. W. Lau, R. W. L. Ip, H.K. Chan, S. Kong. "Implementation of total productive maintenance: A case study.”. International Journal of Production Economics. 2005, 95.1: 71-94.

[2] Lee, J., E., Lapira, B., Bagheri, H. A., Kao. Recent advances and trends in predictive manufacturing systems in big data environment”. Manufacturing Letters, 2013, 1.1: 38-41.

[3] Shi, J., J., Wan, H., Yan, H. Suo. "A survey of cyber-physical systems." Wireless Communications and Signal Processing (WCSP), 2011 International Conference on. IEEE, 2011.

[4] Kopetz, H. Internet of things. In: Real-time systems (pp. 307-323). Springer US. 2011.

[5] Lee, J., B., Bagheri, H. A., Kao. "A cyber-physical systems architecture for industry 4.0-based manufacturing systems." Manufacturing Letters, 2015, 3: 18-23.

[6] Vijaya Raghavan, A., W., Sobel, A., Fox, D., Dornfeld, P., Warnhoff. “Improving machine tool interoperability using standardized interface protocols: MT Connect.” In: Proceedings of the 2008 international symposium on flexible automation (ISFA), 2008, Atlanta, GA, USA.

[7] Yang, S., B., Bagheri, H. A., Kao, J., Lee. "A Unified Framework and Platform for Designing of Cloud-based Machine Health Monitoring and Manufacturing Systems.", Journal of Manufacturing Science and Engineering, 2015.

## AUTHOR

My Name is **Temitope Awodiji**, and I work remotely as a Data Analyst. I hold a master’s degree in Computer Information Science. I am an Efficient Data Analyst professional with expert skills in SQL, Power BI, Tableau, EXCEL, and other data analytics tools. My experience includes generating, manipulating, interpreting, and analyzing data in fast-paced delivery and operations. Growing up, I have always enjoyed solving puzzles. So, this is the same way I see Data Set. I see it as a puzzle I want to solve. Finding the patterns nobody sees is a challenge to me.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# PREDICTION OF VACCINATION SIDE-EFFECTS USING DEEP LEARNING

Farhan Uz Zaman, Tanvinur Rahman Siam and Zulker Nayen

Department of Computer Science and Engineering,
BRAC University, Dhaka, Bangladesh

## *ABSTRACT*

*Deep learning has been very successful in the field of research which includes predictions. In this paper, one such prediction is discussed which can help to implement safe vaccination. Vaccination is very important in order to fight viral diseases such as covid-19. However, people at times have to go through unwanted side effects of the vaccinations which might often cause serious illness. Therefore, modern techniques are to be utilised for safe implementations of vaccines. In this research, Gated Recurrent Unit, GRU, which is a form of Recurrent Neural Network is used to predict whether a particular vaccine will have any side effect on a particular patient. The extracted predictions might be used before deciding whether a vaccine should be injected to a particular person or not.*

## *KEYWORDS*

*Deep Learning, Gated Recurrent Unit, Recurrent Neural Network.*

## 1. INTRODUCTION

Viral diseases have taken millions of lives over the years. Spanish Flu and Covid-19 have proven to be the darkest enemies of humanity. Not only the loss of lives, these viruses have also caused huge economic losses. Hence, the people who survived the pandemic had to face economic hardships and even famine. Vaccination is very important especially for fighting against pandemics such a covid-19. A national public health institute in the United States of America, Centers for Disease Control and Prevention (CDC), have reported that 23.3 million people have been saved worldwide since 2011 by vaccines [1]. However, the entire process of vaccination is very delicate since human life is more fragile than we think. Therefore, a single mistake might be very costly in terms of life as we have seen in the Cutter Incident back in 1955 which ended with the death of 10 children [2]. To prevent such accidents from occurring several measures have been taken. Since many delicate problems have been solved with the help of modern technologies, turning to deep learning for the prevention of such mishaps would be very ideal.

Deep learning methods such as recurrent neural networks have proved to be very efficient in terms of deriving the correct result within the shortest period of time. In most cases, they are used for the purpose of prediction. Hence, in this paper, a model is developed using a gated recurrent unit (GRU), which is a very popular form of recurrent neural network. This model is used to predict whether a certain vaccine will have any side effect on a person. This model is necessary since administering vaccines safely is a very important issue. Despite the fact that GRU is reputed to have very high accuracy while solving problems, our topic was very challenging since gathering the news of side effects are often buried by the vaccine producers to prevent the

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 295-304, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111424

organisation from getting defamed. Also, because the adverse effect of a vaccine is dependent on a number of factors, including the genetic background of the person receiving the vaccine.

In this paper, the previous works in this field has been discussed in section 2. In section 3, we have explained our model that predicts the side effects of vaccines on individuals. In that section we discussed about the data that has been collected and used in this research and the variables that we included in the data to train and test the model. In the next section, section 4, we shared about the architecture that has been used. The results are shown with respect to accuracy. In section 5, future plans with this model and conclusions in section 6 has been discussed.

## 2. BACKGROUND

An outbreak model with the rate of nonlinear vaccination is being studied. To stop diseases various kinds of vaccination have created this vaccination to stop diseases to spread and save human lives but, in some cases, we also found that this vaccination fails to do its job moreover it creates die effect or death to the who takes the vaccine. Mankind came across various kinds of pandemics. The edge that chooses whether or not the infection ceases to exist is found. The infection free harmony around the world is stable under the limit.

The disease-free balance is unstable above the threshold, and the endemic equilibrium occurs and is asymptotically stable locally. In some particular cases, the global stability of the endemic equilibrium is demonstrated by the limiting theory and the work of the Lyapunov function. Finally, for two special cases of the rate of infection, some numerical simulations are given [3].

Immunization assumes a fundamental part in annihilating various lethal illnesses everywhere in the world. Youth immunization is a demonstrated device for controlling irresistible illnesses and is acknowledged wherever to be protected. The expanded Program on Immunization (EPI) was set up by the World Health Organization (WHO) in 1977 to give general vaccination for all youngsters by 1990. In Bangladesh, EPI was started in 1979, and this is one of the extremely effective programs in the Bangladesh wellbeing area with an inclusion rate of around 85%. Immunizations are given locally as indicated by a yearly timetable in Bangladesh which is known as miniature arrangement. At present, the inoculation plan (miniature arrangement) is created physically including around 5000 Unions and 500 Upazilas, which are regulatory units in Bangladesh. Consequently, ongoing data isn't accessible at the focal office for checking reasons. In any case, for additional fortifying of the inclusion and compelling checking, accessibility of data is pivotal. To address this issue, the paper proposes a computerized interaction for building up the inoculation plan program through a Web-based MIS programming application. The people group well-being laborers and authorities will be ready to get to the framework through their versatile remote gadgets effectively accessible to them. The timetable and staff data will be accessible to the mid-level and focal authorities which will guarantee legitimate observing, responsibility, straightforwardness simultaneously. Subsequently, the inclusion of immunized youngsters in Bangladesh will be improved through adjusting this arrangement by the public authority [4].

Furthermore, in another paper [5], the writer sums up and dissects the stochastic Hepatitis B pandemic model with amazing inoculation and changes with massive population growth. Hepatitis B virus is one of the most serious global health issues, as we all know. Hepatitis B virus is spread through physical touch, sexual contact, by taking polluted blood or the baby can also be affected by the mother during pregnancy. From the report from WHO every year massive amounts are getting infected by Hepatitis B virus and around 600000 deaths are being caused by Hepatitis B virus. To fight over this risk, they proposed that mathematical models will be more efficient. As per their analysis, mathematical models will give more success rate. So, they started working in utilizing the stochastic differential condition with Levy leaps to consider the asymptotic conduct of the Hepatitis B pestilence model. Considering the aggregated bounce size, the edge of our plague model is explored, which decides the perseverance or annihilation of Hepatitis B. Mathematical reproductions are acquainted with the legitimacy of our outcomes.

Irresistible illness is one of the medical problems that compromise the total populace. The spread of irresistible illnesses can be controlled utilizing immunization projects and a self-detachment as proposed by the general well-being association [6] [7]. The people's motivators to take the control measures rely upon their dreadfulness of the illness. This frightfulness or a sensation of dread is related to human conduct. Thus, the goal of this study is to fuse human conduct in an infection demonstrating. Intending to contemplate the degree of dreadfulness in both sub-populations of accepting total data and getting deficient data and its effect on the immunization program. In this paper, a model is created by utilizing individual-based displays to deal with catching the human social changes during a sickness episode. The investigation of the outcomes introduced the connection between frightfulness and the inoculation of the choice of the person [8]. Henceforth, the degree of frightfulness should be inspected to sorts out a powerful inoculation program.

In another paper [9], the writers appointed out that, taking a vaccine is one of the best general wellbeing mediations, because of its security and adequacy, however, inoculation doesn't generally mean vaccination. Various angles related both to the person that gets the antibody and the explicitness of every immunization controlled are important for the way toward acquiring satisfactory vaccination, and it is fundamental to notice the perspectives to stay away from antibody disappointments. The examination of immunogenicity and adequacy reads for the measles, varicella, and mumps immunizations highlight the need to join two portions into the fundamental inoculation schedules to control these infections [10]. Epidemiological investigations that examined flare-ups of these infections distinguished cases in people that got two portions of the antibody, which may demonstrate likely auxiliary disappointment. For the yellow fever antibody, the current conversation lies in the ideal number of dosages for singular security. Notwithstanding a couple of reports in the writing concerning antibody disappointments, immunogenicity contemplates exhibit winding down security throughout the long term, predominantly in the paediatric age section [11]. In the flow situation of end and control of sicknesses, related with the abatement in the dissemination of the wild-type infections, the job of epidemiological reconnaissance is significant for extending information on the numerous variables included, coming full circle in immunization disappointments and the development of episodes. Flare-ups of antibody-preventable sicknesses contrarily sway the believability of inoculation programs, prompting low immunization inclusion rates and meddling in inoculation's prosperity.

## 3. MODEL AND IMPLEMENTATION

The model we proposed to detect side effects of vaccines consists of simple but effective steps. Below is a diagram of the proposed model that consists of the steps to be followed in the attempt to detect the side effect of vaccines.

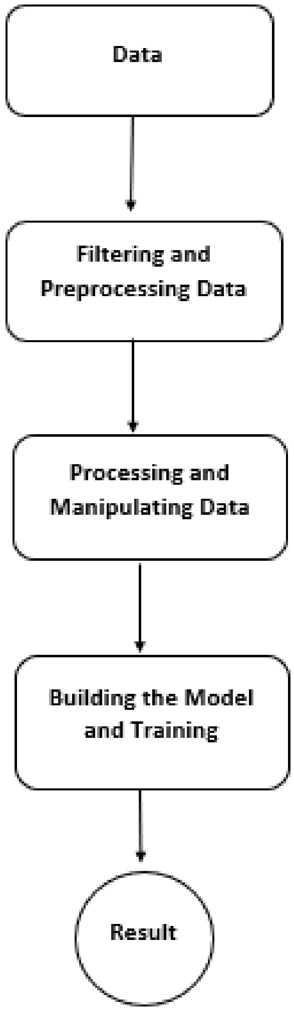
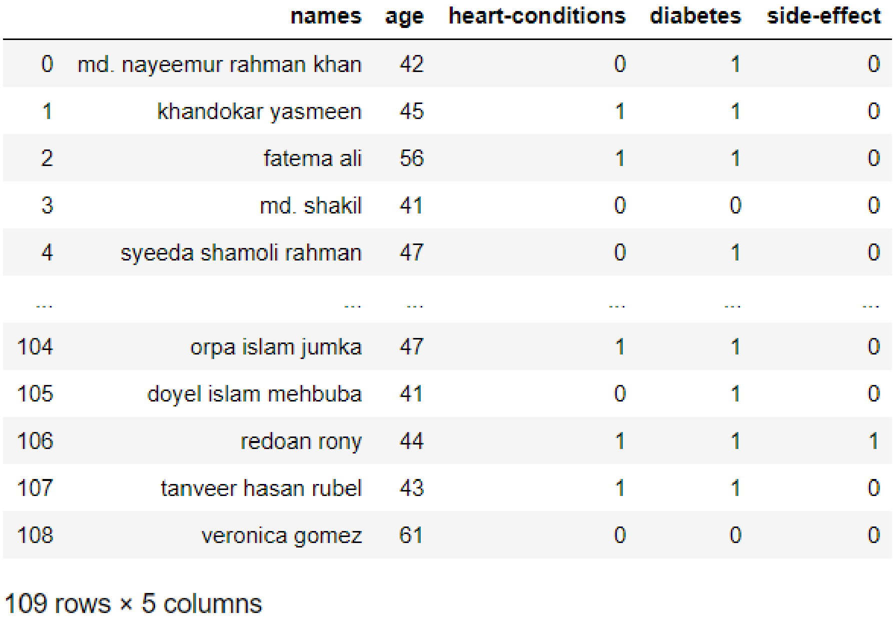

Data
Filtering and Preprocessing Data
Processing and Manipulating Data
Building the Model and Training
Result

Figure 3.1. Workflow of System

At the very beginning, we collect data from several centers in the city where vaccines are being administered. Because of the movement limitations and other restrictions imposed during the pandemic for the sake of safety and preventing the virus from spreading, the data acquired is relatively limited. The data is then pre-processed, processed and manipulated according to our needs. The data we used consisted of two variables. The variables that we used for this model are Diabetes and Cardiovascular Diseases. These two variables were used because it is prominent among the population of the world. So, more data was available that included these factors. Furthermore, these variables have proven to have the most significant effects on people who had vaccines administered. However, these aren't the only factors that can influence vaccine effectiveness. Below we have provided a glimpse of the data that we have collected and used to this research.



| | names | age | heart-conditions | diabetes | side-effect |
|---|---|---|---|---|---|
| 0 | md. nayeemur rahman khan | 42 | 0 | 1 | 0 |
| 1 | khandokar yasmeen | 45 | 1 | 1 | 0 |
| 2 | fatema ali | 56 | 1 | 1 | 0 |
| 3 | md. shakil | 41 | 0 | 0 | 0 |
| 4 | syeeda shamoli rahman | 47 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... |
| 104 | orpa islam jumka | 47 | 1 | 1 | 0 |
| 105 | doyel islam mehbuba | 41 | 0 | 1 | 0 |
| 106 | redoan rony | 44 | 1 | 1 | 1 |
| 107 | tanveer hasan rubel | 43 | 1 | 1 | 0 |
| 108 | veronica gomez | 61 | 0 | 0 | 0 |

109 rows × 5 columns

Figure 3.2. Glimpse of the dataset.

As we have mentioned above, two very serious variables for the side effects are being considered in the paper, which are diabetes and cardiovascular diseases. We have found that a large portion of the people who have been administered the vaccines have diabetes. Below we have provided a bar chart generated from the data we have gathered comparing the number of people who have diabetes to those who do not, from the people who got the vaccines administered.

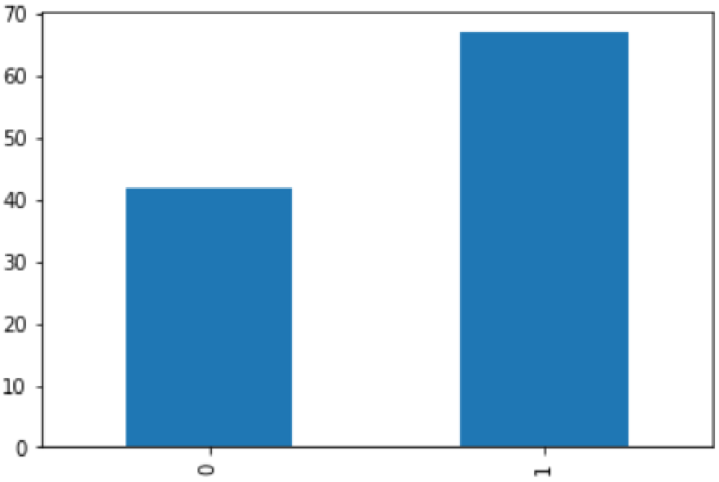


Figure 3.3. Comparison between number of people having and not having Diabetes.

We also discovered that half of those who received the vaccinations had cardiovascular illness. Another bar chart has been created that compares the number of people with cardiac issues to those who have not, among those who have received immunizations.

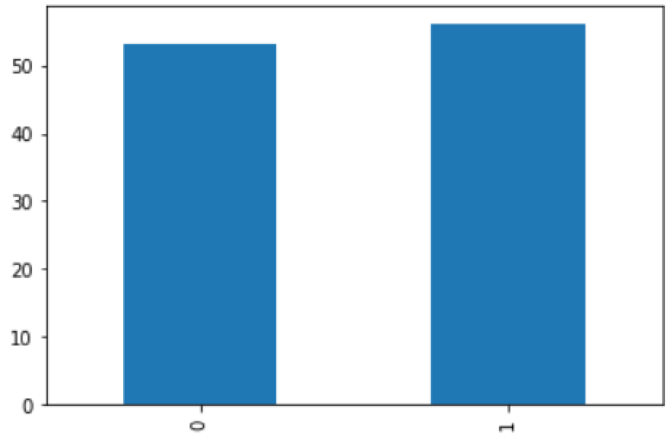


Figure 3.4. Comparison between number of people having and not having Heart Disease.

This is mainly because at the time of the data collection only the elderly citizens were being vaccinated, who were prone to either one or both the conditions. And very less people had none of the conditions. Below we have generated another bar chart comparing the number of people have any of the two conditions to those who have none.

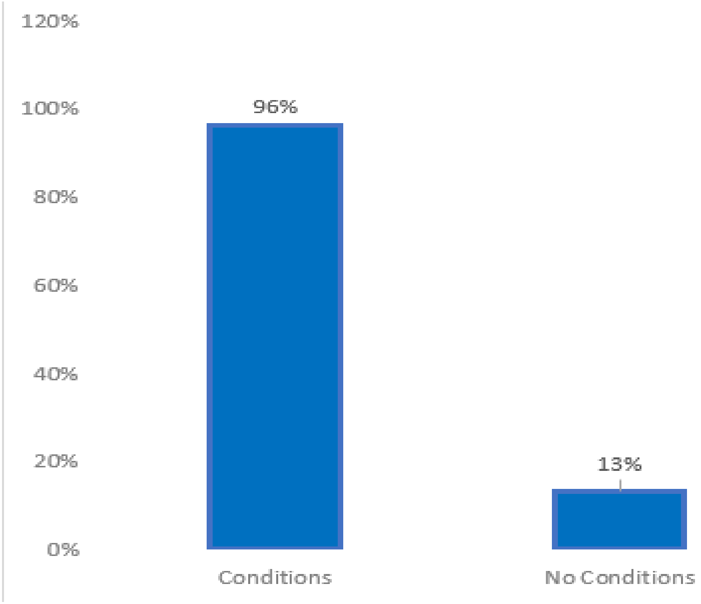


Figure 3.5. Comparison between number of people having condition and who having none.

Finally, we looked at how many persons experienced side effects or other forms of discomfort after receiving the vaccines.

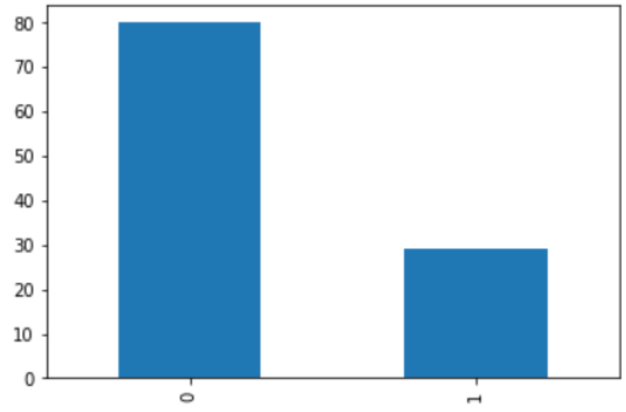
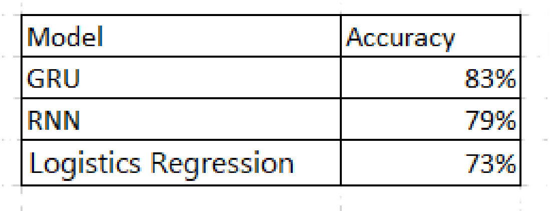


Figure 3.6. Comparison between number of people having and not having Side-effects.

## 4. Results

The processed data is then put inside an Artificial Neural Network model. Several models of ANN have been tested to find the one that gives us the optimum accuracy with the given set of data. Recurrent neural networks such as LSTM [12] and GRU showed results with higher accuracy. The GRU model was then selected due to having the highest accuracy. The GRU and the other models were trained with 80 percent of the data and tested with the remaining 20 percent. The models predict whether a vaccine when administered will have any effect on the person or not. The result of the trained ANN is given in the table below.

Table 4.1. Model vs Accuracy



| Model | Accuracy |
|---|---|
| GRU | 83% |
| RNN | 79% |
| Logistics Regression | 73% |

The graphical form of the table is being shown for better visualisation of the accuracy of each of the models which were trained for this research.

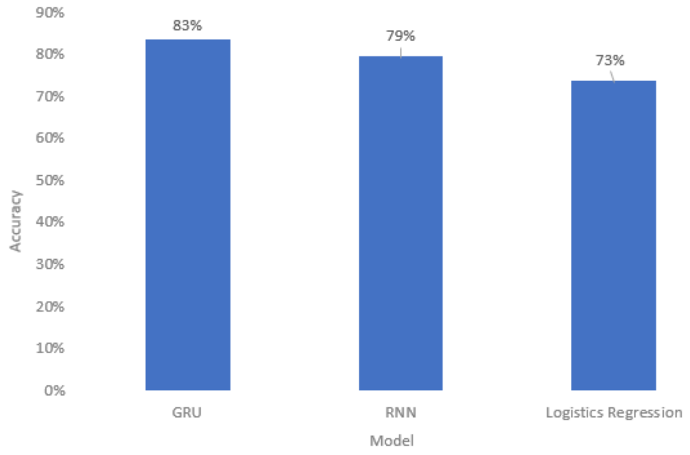


Figure 4.1. Graph of Model vs Accuracy

The GRU model is a small Artificial Neural Network model from the Recurrent Neural Network family which is very similar to that of the LSTM architecture. The details of only this model is explained below since we are focusing on the model with the highest accuracy.

This model is a linear stack of layers and has been built with an instance of the sequential class. The first layer of this model is the embedding layer which receives as input an integer matrix of size (32,50). This layer gives an output of shape (*, 50, 32). A GRU layer has been added followed by this. The GRU layer has 100 units, which proved to be sufficient enough for the purpose of producing a better accuracy than the LSTM layer. The unit of the GRU layer was determined on a trial-and-error basis. The final layer added to this small but efficient model was the dense layer. The layer had a sigmoid activation argument passed to it. The output from this layer is an array of shape (*,6). This model had no dropout layer unlike most LSTM models. [13].

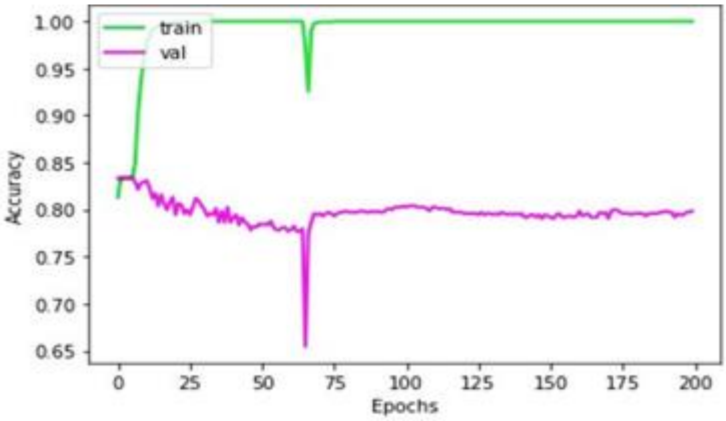


Figure 4.2. Graphical View of Accuracy

## 5. FUTURE PLAN

Through this paper we tried establishing the point that neural networks can be used to predict ill effects of vaccines on individuals. Our goal for the future is to collect more first-hand data, if possible, which would reduce the amount of over-fitting that might have happened during this research, and to include more variables that would make the model more reliable. It is expected that as the vaccine roll-out ramps up in different country, we would be able to differentiate the different demographic variables, which would enhance the research. We would also like to try out different variations of neural networks to figure out which one would serve our purpose the best. For better research purposes we could also segment the data based on area, ethnicity, etc. This would not only make healthcare more advanced but also ensure the safe administration of vaccines, helping us to fight against different lethal microbes in the future.

## 6. CONCLUSIONS

Predicting whether a vaccine would have any particular side effect when administered on a person is very important since there are a lot of viruses which present the vulnerability of mankind and its existence. This will not only help us to achieve the trust of the mass population who are scared of side effects but will also move towards a more immune world. However, this is one of the very first models to predict such a thing. This model needs to be redefined over the time so achieve better accuracy with more variables included, since there are a lot of variables that need to be kept in mind. As we have mentioned above, in this model we only took two variables into account, diabetes and CVD, but in a real-world scenario a person maybe exposed to other health issues as well. We hope our research would bring out the best possible results for the achieving the herd immunity which is very vital for out battles against pandemics which we have faced time and again.

## ACKNOWLEDGEMENTS

At the beginning, we thank Allah because of His blessings, which has enabled us to continue our research without facing any major difficulties. In addition, we wanted to thank all the supportive faculty members and our supervisor in particular for tolerating our mistakes and providing continuous input to enhance our research. Moreover, we thank our parents as well as team-mates who have given us tremendous support.

## REFERENCES

[1] "Vaccine Crisis first polio vaccine", https://www.ncbi.nim.gov/pmc/articles/PMC1383764/, accessed: 2020-11-05.

[2] "Vaccines Failed the national interest", https://nationalinterest.org/blog/coronavirus/four-times-history-vaccinesfailed-lessons-coronavirus-vaccine-166116, accessed: 2021-01-03.

[3] Y.-L. Yang, J.-Q. Li, and J.-G. Zhang, "Global analysis for an epidemic model with nonlinear vaccination rate," in *2009 International Conference on Machine Learning and Cybernetics*, vol. 4. IEEE, 2009, pp. 2096–2100.

[4] F. Nadi, "2018 5th asia-pacific world congress on computer science and engineering (apwc on cse 2018)," 2018.

[5] D. Kiouach and Y. Sabbar, "Threshold analysis of the stochastic hepatitis b epidemic model with successful vaccination and levy jumps," in *2019 4th World Conference on Complex Systems (WCCS)*. IEEE, 2019, pp. 1–6.

[6] T. S. Li, J. Labadin, P. Piau, S. Abd Rahman, and L. Y. Tyng, "The effect of vaccination decision in disease modelling through simulation," in *2015 9th International Conference on IT in Asia (CITA)*. IEEE, 2015, pp. 1–6.

[7] J. Marlet, C. Gaudy-Graffin, D. Marc, R. Boennec, and A. Goudeau, "Factors associated with influenza vaccination failure and severe disease in a french region in 2015," *Plos one*, vol. 13, no. 4, p. e0195611, 2018.

[8] E. L. Pesanti, "Immunologic defects and vaccination in patients with chronic renal failure," *Infectious disease clinics of North America*, vol. 15, no. 3, pp. 813–832, 2001.

[9] T. C. d. M. B. Petraglia, P. M. C. d. M. Farias, E. M. d. Santos, D. A. d. Conceic¸ao, M. d. L. d. S. Maia˜ *et al.*, "Vaccine failures: assessing yellow fever, measles, varicella, and mumps vaccines," *Cadernos de Saude P´ ublica´* , vol. 36, p. e00008520, 2020.

[10] M. K. Patel and W. A. Orenstein, "Classification of global measles cases in 2013–17 as due to policy or vaccination failure: a retrospective review of global surveillance data," *The Lancet Global Health*, vol. 7, no. 3, pp. e313–e320, 2019.

[11] H. Chisholm, A. Howe, E. Best, and H. Petousis-Harris, "Pertussis vaccination failure in the new zealand pediatric population: Study protocol," *Vaccines*, vol. 7, no. 3, p. 65, 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

## AUTHORS

**Farhan Uz Zaman** was born in Dhaka, Bangladesh. He graduated from the department of Computer Science and Engineering of a very reputed private university of Bangladesh, BRAC University. He is a Machine Learning (ML) and Deep Learning (DL) enthusiast, having several works with Scikit Learn and TensorFlow where the most popular DL models such as CNN and RNN are his favourite things to play with.

**Zulker Nayen** was born in Feni, Bangladesh. He is studying at BRAC University in the department of Computer Science and Engineering. He is a Machine Learning (ML) and Bioinformatics enthusiast. He is doing internship at Desh Cyber Ltd and doing some projects for Computer Club of BRAC University.

**Tanvinur Rahman Siam** was born in Dhaka, Bangladesh. He is studying at BRAC University in the department of Computer Science and Engineering. He is a Blockchain and Deep Learning (DL) enthusiast. He has done internship in Dhaka Tribune and he loves to participate in programming contest.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# END-TO-END CHINESE DIALECT DISCRIMINATION WITH SELF-ATTENTION

Yangjie Dan, Fan Xu*, Mingwen Wang

School of Computer Information Engineering,
Jiangxi Normal University, Nanchang 330022, China

## *ABSTRACT*

*Dialect discrimination has an important practical significance for protecting inheritance of dialects. The traditional dialect discrimination methods pay much attention to the underlying acoustic features, and ignore the meaning of the pronunciation itself, resulting in low performance. This paper systematically explores the validity of the pronunciation features of dialect speech composed of phoneme sequence information for dialect discrimination, and designs an end-to-end dialect discrimination model based on the multi-head self-attention mechanism. Specifically, we first adopt the residual convolution neural network and the multi-head self-attention mechanism to effectively extract the phoneme sequence features unique to different dialects to compose the novel phonetic features. Then, we perform dialect discrimination based on the extracted phonetic features using the self-attention mechanism and bi-directional long short-term memory networks. The experimental results on the large-scale benchmark 10-way Chinese dialect corpus released by IFLYTEK[1] show that our model outperforms the state-of-the-art alternatives by large margin.*

## *KEYWORDS*

*Dialect discrimination, Multi-head attention mechanism, Phonetic sequence, Connectionist temporal classification.*

## 1. INTRODUCTION

With the gradual advancement and promotion of Putonghua, the hometown dialects of many provinces and cities in China have been gradually assimilated by Putonghua. According to statistics from United Nations Educational Scientific and Cultural Organization (UNESCO), a language will disappear every two weeks. However, Chinese dialect, as an excellent intangible cultural heritage of the Chinese nation, should not disappear with the popularization of Mandarin. As a special language variant, Chinese dialect has always been a research hotspot in linguistics. It is urgent to protect dialects.

In 2018, based on the "Dialect Protection Plan", iFLYTEK released the world's first large-scale 10-way precious dialect (Ningxia, Hefei, Sichuan, etc.) phonetic and phoneme corpora covering most parts of my country in order to jointly advance the algorithm research and protection of dialects. Traditional language recognition methods[2] focus on the underlying acoustic features,

* Corresponding author: xufan@jxnu.edu.cn

[1] https://www.iflytek.com/index.html

[2] Language recognition task learns the distinguishing characteristics between different languages through speech sentences and corresponding language tags; dialect recognition is a special case in language recognition. The task of identifying dialect types is to distinguish different language variants in the same language. Compared with distinguishing different languages, distinguishing dialects is more challenging.

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 305-316, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111425

such as MFCC (mel-frequency cepstral coefficients) and Fbank (log mel-filterbank), without considering the meaning of the pronunciation itself, resulting in poor performance. In fact, when human beings distinguish different types of dialects, they often judge them by the pronunciation characteristics of the dialect itself. More specifically, we investigated the pronunciation dictionary[3] of Chinese dialects to list the phoneme forms of "fang" and "yan" in Minnan dialect, Guangzhou dialect, Hakka dialect and Shanghai dialect as shown in Table 1. It can be seen from Table 1 that the corresponding pronunciation forms (phonemes) of the same Chinese characters in different dialects are completely different. In other words, if we can effectively extract the unique pronunciation features of different dialects, we can use the pronunciation features of dialects to better distinguish different types of dialects.

Table 1. Examples of phonemes in different Chinese dialects.

| Dialect type | fang | yan |
|---|---|---|
| Minnan dialect | beng1 hng1 hong1 | ngian2 |
| Guangzhou dialect | fong1 | jin4 |
| Hakka dialect | fong1 | ngien2 |
| Shanghai dialect | faon | gni re yi |

Based on this observation, this paper systematically explores the effectiveness of the dialect phonetic features composed of phoneme sequence information for language discrimination, and designs an end-to-end dialect discrimination model based on the multi-head self-attention mechanism. The model first adopts residual CNN (convolutional neural networks)[11] and multi-head attention mechanism to effectively extract the unique phoneme sequence information of different dialects to generate voice pronunciation features, and then uses attention mechanism and bidirectional long short-term memory (BiLSTM)[18] for dialect discrimination. We conducted experiments on the large-scale benchmark 10-way dialect corpus released by iFlytek. The experimental results show that the multi-headed self-attention mechanism [30] can effectively extract the pronunciation characteristics of phoneme sequences unique to different dialects, which greatly improves the discrimination performance of dialects.

The follow-up content of this paper is organized as follows: Section 2 presents the related work of language discrimination in recent years; Section 3 illustrates our model in detail; Section 4 introduces the data set, experimental settings and detailed analysis of experimental results; Section 5 concludes the paper.

## 2. RELATED WORK

This section mainly introduces representative language discrimination models from two perspectives: traditional acoustic features based and speech pronunciation features based models.

### 2.1. Methods based on Traditional Acoustic Features

Traditional language discrimination methods adopt underlying acoustic features to build acoustic models in order to obtain fixed encoding vectors of speech sentences. At present, the commonly used artificially extracted underlying acoustic features include: Fbank (log mel-filterbank), MFCC (mel-frequency cepstral coefficients), PLP (perceptual linear prediction), SDC (delta coefficients) [1,2], etc. Since the underlying acoustic features are extracted in units of frames, the number of frames corresponding to speech sentences with different durations is also different.

[3] http://cn.voicedic.com/

Therefore, how to convert a variable-length speech sentence into a fixed vector representation is a vital step. The typical methods are GMM (Gaussian mixture model) super vector [3] and GMM i-vector [4]. The i-vector feature contains relevant information about the speaker and language. This feature is usually used as a speech sentence representation to train a language classifier. Commonly used classifiers include multi-class logistic regression and support vector machines. But the main disadvantage of the i-vector method is its poor discrimination effect on short speech sentences [5].

Recently, as deep learning technology has achieved great success in speech recognition tasks[6], some researchers have begun to explore language discrimination technology based on deep learning. In the early days, many studies [7,8,9] used deep learning technology to extract the bottleneck features of speech sentences, and achieved better language discrimination performance. Currently, some researchers recognize the powerful representation capabilities of deep learning and directly use various types of neural networks to build end-to-end language discrimination models. It was first used by Lopez-Moreno et al.[5] to successfully use deep neural networks for language discrimination. The network directly takes the underlying acoustic features of the speech sentence, and then scores each frame on different languages. The score of the speech sentence is the average of all frames within the sentence. After that, there are many language discrimination models with different structures, such as deep neural networks (DNNs) based on attention mechanism [10], convolutional neural networks (CNNs) [11,12,13,14], delayed neural networks [15,16] and recurrent neural networks (RNNs). Because the RNN network has a strong ability to extract context-related (global) features, it can learn better feature representations for the temporal feature characteristics of speech, which improves the performance of the language discrimination. In practical, there are several different variants of recurrent neural networks, including gated recurrent unit recurrent neural network (GRU) [17], long and short-term memory recurrent neural network [18,19,20,21,22], Bidirectional long and short-term memory recurrent neural network [23,24].

## 2.2. Method Based on Pronunciation Characteristics

Traditional language discrimination models based on underlying acoustic features ignore many important speech pronunciation information. Therefore, Tang et al. [25] adopted the pronunciation characteristics of speech to improve the effect of language discrimination. The specific method was to use a speech phoneme recognition model to extract the frame-level speech pronunciation characteristics, and then feed the speech pronunciation characteristics into the language discrimination model. The speech phoneme recognition model of this method uses a cross-entropy loss function. Recently, studies have shown that the end-to-end acoustic model based on CTC [26] (connectionist temporal classification) has obtained better performance [27, 28, 29].

Based on these observation, this paper proposes an end-to-end dialect discrimination model based on multi-head self-attention mechanism. We adopt the CTC loss function to train the acoustic model of speech phoneme recognition, and integrate a multi-headed self-attention layer, which can give the acoustic model the ability to extract unique pronunciation characteristics of different dialects. The multi-head self-attention mechanism is based on the transformer [30] model proposed by Google in 2017. This model performs very well on machine translation, and it also has a good effect on speech recognition [31] tasks.

# 3. SELF-ATTENTION DRIVEN DIALECT DISCRIMINATION

Figure 1 illustrates the proposed dialect discrimination model. The model is mainly composed of two parts. The top side of the figure is the speech phoneme recognition model, and the main function of this part is to extract the pronunciation characteristics of the dialect. The bottom side of the figure is the dialect discrimination model, which mainly uses the pronunciation characteristics of dialects to improve the accuracy of dialect discrimination. In the speech phoneme recognition model, we extract more abstract local features of the speech through a residual CNN, and then feed them into a multi-headed self-attention layer, which can pay attention to the relationship between each frame of speech and other frames, and then map to the appropriate dimension through a fully connected layer, and finally calculate the difference between the predicted phoneme sequence and the real phoneme sequence through the CTC loss function. When identifying dialect types, we designed two models. One is to input the recognized dialect pronunciation features into Self-Attention Pooling (SAP) [34]. This attention mechanism can encode the variable-length dialect pronunciation features into a fixed vector representation, which is then input to the fully connected layer. In fact, we can also do an average pooling or maximum pooling of pronunciation features, but the attention mechanism is essentially a weighted average operation on the pronunciation features, and the weighted average can include these two situations according to the different proportions of the allocation. Then, we feed the recognized pronunciation features of dialects into the BiLSTM network. We use the output of the last moment of BiLSTM as the fixed-length vector representation of the speech sentence, then it is mapped to 10 dialects through two fully connected layers. The first fully connected layer maps the input features to each implicit semantic node, and the second fully connected layer represents the display expression of the classification. Finally, the probability of the speech sentence belonging to each dialect is obtained through softmax. The function of each sub-module is introduced below.

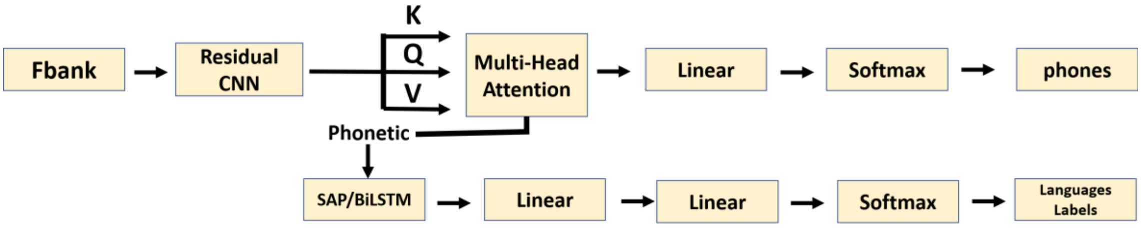


Figure 1. Dialect discrimination model based on self-attention mechanism

## 3.1. Residual CNN

Residual network [32] was first applied to image classification, and Li et al. [33] adopted residual CNN to extract speech features and performed language discrimination. In fact, CNN can better extract features on voice frequency, and residual CNN can use a deeper network to extract more abstract voice features. Due to the existence of the residual mechanism, even if the number of network layers increases, it will not cause network degradation. In order to obtain a more abstract representation of the speech sentence, we design the residual CNN network structure based on resnet18.

## 3.2. Self-Attention Mechanism

The self-attention mechanism is a coding sequence scheme proposed by the Google team Vaswani et al. [30] in 2017. It can be considered that it is a sequence coding layer like general

CNN and RNN. The self-attention mechanism is a special attention mechanism, and it only needs a separate sequence to calculate the code of this sequence. The self-attention mechanism uses standard dot product attention, and its calculated attention weight is shown in formula (1):

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{1}$$

where the dimensions of the query vector $Q$ and the key vector $K$ are both $d_k$, and the length of the value vector $V$ is $d_v$.

The multi-head self-attention mechanism is adopted to calculate a single attention multiple times, but the query vector, key vector, and value vector are different each time. Specifically, the multi-head self-attention mechanism layer first generates $h$ different query vectors $Q$, key vectors $K$, and value vectors $V$, where the dimensions of the query vector and the key vector are $d_k$, and the dimension of the value vector is $d_v$. For each set of query vectors, key vectors and value vectors, a vector with dimension $d_v$ can be generated by formula (1), and then the generated $h$ vectors can be spliced together. The above process can be described by formula (2) and formula (3):

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(head_1, \ldots, head_h)\boldsymbol{W}^{\boldsymbol{O}} \tag{2}$$

$$head_i = Attention(\boldsymbol{QW}_i^{\boldsymbol{Q}}, \boldsymbol{KW}_i^{\boldsymbol{K}}, \boldsymbol{VW}_i^{\boldsymbol{V}}) \tag{3}$$

where $\boldsymbol{W}_i^{\boldsymbol{Q}} \epsilon R^{d_{model} \times d_k}$, $\boldsymbol{W}_i^{\boldsymbol{K}} \epsilon R^{d_{model} \times d_k}$, $\boldsymbol{W}_i^{\boldsymbol{V}} \epsilon R^{d_{model} \times d_v}$, $\boldsymbol{W}_i^{\boldsymbol{O}} \epsilon R^{hd_v \times d_{model}}$, $h$ is the number of heads, $d_{model}$ is the model dimension.

## 3.3. Connectionist Temporal Classification (CTC)

The CTC loss function can select the sequence with the greatest probability from the given input sequence [26]. We use $x$ to represent the input sequence and $z$ to represent the corresponding phoneme sequence. Each training sample can be represented by a tuple (x, z). Our goal of maximizing the likelihood function is to minimize the value of formula (4).

$$\mathcal{L}_{ctc} = -ln \prod_{(\boldsymbol{x},\boldsymbol{z})\epsilon S} p(\boldsymbol{z}|\boldsymbol{x}) = -\sum_{(\boldsymbol{x},\boldsymbol{z})\epsilon S} lnp(\boldsymbol{z}|\boldsymbol{x}) \tag{4}$$

where $(x, z) \in s$ represents a training sample.

## 3.4. Bidirectional Long Short-Term Memory Network

LSTM is a special recurrent neural network, which has the ability to learn long-term dependence, and is suitable for processing and predicting important events with relatively long intervals and delays in time series. The relevant parameter calculation formula of the LSTM model is as follows:

$$\boldsymbol{i}_i = sigmoid(\boldsymbol{W}_i[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_i) \tag{5}$$

$$\tilde{\boldsymbol{c}_i} = \tanh(\boldsymbol{W}_c[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_c) \tag{6}$$

$$\boldsymbol{f}_i = sigmoid(\boldsymbol{W}_f[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_f) \tag{7}$$

$$\boldsymbol{o}_i = sigmoid(\boldsymbol{W}_o[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_o) \tag{8}$$

$$\boldsymbol{c}_i = \boldsymbol{i}_i * \tilde{\boldsymbol{c}_i} + \boldsymbol{f}_i * \boldsymbol{c}_{i-1} \tag{9}$$

$$\boldsymbol{h}_i = \boldsymbol{o}_i * tanh(\boldsymbol{c}_i) \tag{10}$$

From the parameter calculation formula of the LSTM model, it is known that the LSTM unit has three "gate" structures that determine the state of the cell. They are input gate, forget gate and output gate. The value of the sigmoid function is between 0 and 1. As the gate structure and the input data are multiplied, the amount of information of the input data can be determined. The input variables that determine the states of these three doors are the same, but the parameters corresponding to the doors of different functions are different. It can be seen from the formula that the states of the three gate structures at a certain moment are all related to the current input $x_i$ and the output value $h_{i-1}$ at the previous moment. The parameters that determine the state of the forget gate are $W_f$ and $b_f$. The state determines the proportion of the previous state information that is forgotten at the current moment. The parameters that determine the state of the output gate are $W_o$ and $b_o$, and the state determines the ratio of the current time and the previous state information. The parameters that determine the state of the input gate are$\boldsymbol{W}_i$ and $\boldsymbol{b}_i$, and this state determines how much information input at the current moment is retained. The BiLSTM takes the order of the sequence into account, that is, there are two layers of one-way LSTM, one layer extracts the global features of the order, and the other layer extracts the global features of the reverse order.

## 4. Experiment and Result Analysis

This section mainly describes the iFLYTEK dialect data set, baseline models, parameter settings, and analysis of experimental results.

### 4.1. Data set

The dialect speech data set used in this experiment was released by iFLYTEK. It contains 10 different dialect speech and phoneme corpora[4]. The statistics of the corpus is shown in Table 2. Each dialect contains an average of 6 hours of reading style speech data, covering 35 people.

Table 2. Statistics of the iFLYTEK dialect data set.

| **Data set** | **Training set** | | | **Development set** | | |
|---|---|---|---|---|---|---|
| Dialect | Speaker | Sentence per person | Total sentences | Speaker | Sentence per person | Total sentences |
| Ningxia dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Hefei dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Sichuan dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Shanxi dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Changsha dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Hebei dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Nanchang dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Shanghai dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Hakka dialect | 30 | 200 | 6000 | 5 | 100 | 500 |
| Minnan dialect | 30 | 200 | 6000 | 5 | 100 | 500 |

[4] http://challenge.xfyun.cn/2018/aicompetition/tech

The data was collected by various types of smart phones, and the recording environment includes a quiet environment and a noisy environment. The data was stored in a PCM format with a sampling rate of 16000 Hz and 16-bit quantization. The data set contains training set and development set. There are 60,000 sentences in the training set, 6000 sentences in each dialect, including 30 speakers, 15 males and 15 females, and 200 voices per speaker; the development set has 5000 voices, each dialect has 500 voices, and each dialect contains 5 speakers, including 2 females and 3 males. Each phonetic sentence also has its corresponding phoneme label, such as "l iou4 sh iii2 _e er4 _v van2 s ii4 f en1". We adopt 60,000 speech sentences as our training set, and take 5000 speech sentences as our testing set.

## 4.2. Baseline Models

We adopt three benchmark models for dialect discrimination. The first one is a dialect discrimination model based on i-vector features, the second is the LSTM-based dialect discrimination model officially provided by iFLYTEK[5], and the third is the single model adopted by the first winner of the first dialect discrimination competition in 2018[6].

**Baseline model 1**: This model proposed a dialect discrimination model based on i-vector features. More specifically, 60-dimensional MFCC features were extracted, which include first-order and second-order difference coefficients. The general background model used to extract i-vector features includes 2048 Gaussian functions, and finally 400-dimensional i-vector features are extracted from each sentence. The baseline model also uses a support vector machine as a classifier.

**Baseline model 2**: This model is officially provided by iFLYTEK. It uses a one-way LSTM and two fully connected layers. The hidden unit of the LSTM has a dimension of 128, and the input dimension of the first layer of fully connected layer is 128. The output dimension is 30. The input dimension of the second fully connected layer is 30, and the final output dimension is 10.

**Baseline model 3**: This model is a single model presented by the first winner of the 2018 first dialect discrimination competition of iFLYTEK. The model is divided into speech phoneme recognition model and dialect discrimination model. The speech phoneme recognition model uses residual CNN and BiLSTM. The dialect discrimination model is to fix the parameters of the trained residual CNN model to remain unchanged, and then add a layer of trainable BiLSTM, and finally integrate the output states of the BiLSTM at all times, that is, use the output state at all times. In short, both the speech phoneme recognition model and the dialect discrimination model have a residual CNN module and their BiLSTM module. For fair comparison, the residual CNN network structure used in this paper is the same as that of this baseline model.

## 4.3. Parameter Settings

We first seperate the original speech sentence into different frames. The window size of the framing is 25ms, and the frame shift is 10ms. Then we use Kaldi[7] toolkit to extract 80-dimensional Fbank features. The relevant parameter settings of residual CNN are shown in Table 3.

[5] http://bbs.xfyun.cn/forum.php?mod=viewthread&tid=39141

[6] http://1024.iflytek.com/h5/?from=singlemessage；The reason why we adopted the first author single system in the competition is that the final system is a composite model, but the composite model is not disclosed. Moreover, 90.50% of the officially announced recognition performance was obtained on the undisclosed final competition test set.

[7] http://www.kaldi-asr.org/

Table 3. Parameter settings in residual CNN network.

| layer | output size | down sample | channels | blocks |
|---|---|---|---|---|
| conv1 | $40 \times L_{in}/2$ | True | 64 | - |
| maxpool | $20 \times L_{in}/4$ | True | 64 | - |
| res1 | $10 \times L_{in}/4$ | False | 64 | 2 |
| res2 | $5 \times L_{in}/4$ | False | 128 | 2 |
| res3 | $3 \times L_{in}/4$ | False | 256 | 1 |
| res4 | $2 \times L_{in}/4$ | False | 512 | 1 |
| avgpool | $1 \times L_{in}/4$ | False | 512 | - |
| reshape | $512 \times L_{in}/4$ | - | - | - |

The parameter settings used by the multi-head self-attention layer are shown in Table 4.

Table 4. Parameter settings in self-attention layer network.

| Model | N | $d_{model}$ | h | $d_k$ | $d_v$ |
|---|---|---|---|---|---|
| Multi-head | 1 | 512 | 8 | 64 | 64 |

where *N* in Table 4 represents the number of layers, and other parameters correspond to formula (2) and formula (3). The hidden state of the BiLSTM is 256 dimensions, and the two-way total has 512 dimensions. BiLSTM is followed by two fully connected layers. The first fully connected layer maps 512 dimensions to 256 dimensions, and the second layer maps 256 dimensions to 10 dimensions. The model in this paper uses the Adam optimization algorithm based on the mini-batch gradient descent algorithm. The optimization algorithm can change the learning rate during the training process and control the step length along the gradient descent through the attenuated learning rate. For the speech phoneme discrimination model, we set a learning rate of 0.0005, and for the language discrimination model, we set a learning rate of 0.001. We train 10,000 frames of speech at the same time each time. This article uses the pytorch framework to implement all network models.

For the evaluation indicators, we use three evaluation indicators to describe the discrimination performance of the system, namely: Accuracy (ACC), Average Decision Cost Function (Cavg) and Equal Error Rate (EER). where the accuracy rate is the evaluation index defined by the IFLYTEK Dialect Competition (the ratio of the number of correct speech sentences to the total number of sentences). Average detection cost and equal error rate are the evaluation indicators used in the standard evaluation of NIST LRE [35].

## 4.4. Result

Table 5 shows the comparison of dialect discrimination performance under various models. It can be seen from the experimental results that the effect of LSTM is slightly worse than that of i-vector, because the discrimination effect of i-vector on short speech (for example, within 3s) is relatively poor [5], but the discrimination effect on relatively long speech is relatively good, and LSTM may not be suitable for processing relatively long speech in the test set [36]. Since BiLSTM can extract context-related features, we use the output vector of the last moment state as a fixed vector representation of a speech sentence. It can be seen in Table 5 that the two models proposed in this paper are better than baseline model 3. In addition, Cavg and EER have also been greatly improved (the smaller the value, the better the performance). Compared with baseline model 3, our model has an extra layer of self-attention. We believe that the self-attention layer can better extract the local part of the speech pronunciation.

Table 5. Performance Comparison of Different Dialect Discrimination Models.

| Model | Acc（%） | $C_{avg}$*100 | EER（%） |
|---|---|---|---|
| i-vector (baseline1) | 74.30 | 9.99 | 10.04 |
| LSTM (baseline2) | 74.28 | 14.03 | 14.76 |
| iFLYTEK 2018 Dialect Competition First List Model (baseline3) | 86.62 | 7.43 | 12.98 |
| Our model (the right side in Figure 1 uses the SAP sub-module) | 87.34 | 6.89 | 5.46 |
| Our model (the BiLSTM sub-module is used on the right in Figure 1) | 89.22 | 5.86 | 4.8 |

Since we adopt the characteristics of dialect pronunciation as the input of the dialect discrimination model, we further compare the phoneme recognition performance of these models. We use a greedy algorithm to decode speech into phoneme sequences. The experimental results are shown in Table 6. WER in the table represents the phoneme error rate. The lower the WER, the better the effect of the speech phoneme recognition model. The WER obtained by our model is higher than the first place system in the iFLYTEK competition. Regarding this phenomenon, we believe that the multi-head self-attention mechanism can better extract the unique pronunciation characteristics of different dialects, and the BiLSTM of the first place system in the iFLYTEK competition is more suitable to extract the pronunciation characteristics commonly shared by different dialects. Therefore, in contrast, the pronunciation features extracted by the multi-head self-attention mechanism are more discriminative in the discrimination of dialect types.

Table 6. Speech phoneme recognition performance comparison.

| Model | WER（%） |
|---|---|
| Residual CNN [32] | 46.57 |
| Our model (after adding multi-head) | 43.08 |
| The phoneme recognition model of iFLYTEK 2018 Dialect Competition No. 1 | 41.06 |

In order to verify the results, we designed a unique dialect phoneme recognition discrimination experiment as shown in Table 7. First, we use the 60,000 phoneme sentences of the training set to train an SVM classifier, which inputs sentence phoneme sequences and outputs dialect types. When testing, we use ASR1 (residual CNN+Multi-Head Attention + CTC) and ASR2 (residual CNN+BiLSTM +CTC), where ASR stands for Automatic Speech Recognition, and the identified phoneme sequence is tested. We first count the unary language models of 10 different dialects in the training set. We believe that the higher the frequency of phonemes in different dialects, the more representative the dialect. We extracted phonemes with word frequency greater than 1%, 0.9%, and 0.8% as features, and we regarded all other phonemes as unregistered words. The word frequency is greater than 1%, 0.9%, and 0.8% have 33, 40 and 47 phonemes, respectively. There are 5000 sentences in the test set.

Table 7. Distinguishing Experiments on Phoneme Recognition of Unique Dialects

| | 1%(33) | 0.9%(40) | 0.8%(47) |
|---|---|---|---|
| **ASR1** | 907 | 936 | 942 |
| **ASR2** | 904 | 925 | 973 |

We found that at 0.9% of the time, 40 phonemes were selected as features, and the recognition effect on ASR1 was better than ASR2, indicating that the multi-headed self-attention mechanism recognized more dialect-specific phonemes.

**CASE STUDY:** Figure 2 and 3 show two instances (example 1 and 2 represented as phoneme respectively; the number 1, 2, 3 and 4 indicate four lexical tones of Chinese).

**Example 1**: m ei2 _i ia1 b u2 sh iii4 _u uo3 z ai4 n a4 n i3 p ei2 _u uo3 l iao2 m a1
**Example 2**: g uo2 j ia1 b o2 _u u4 g uan3 l ao3 d a4 l ao3 d a4

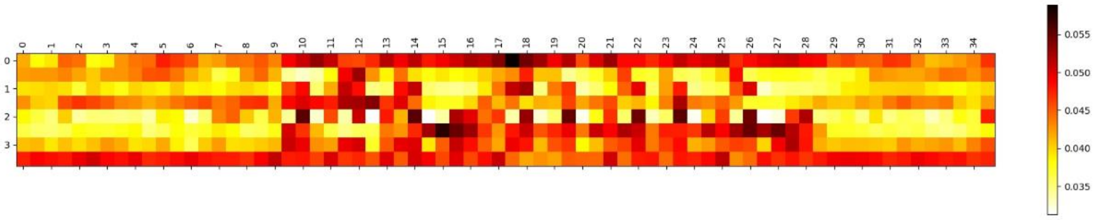


Figure 2. Attention visualization for an instance from shanghai dialect.

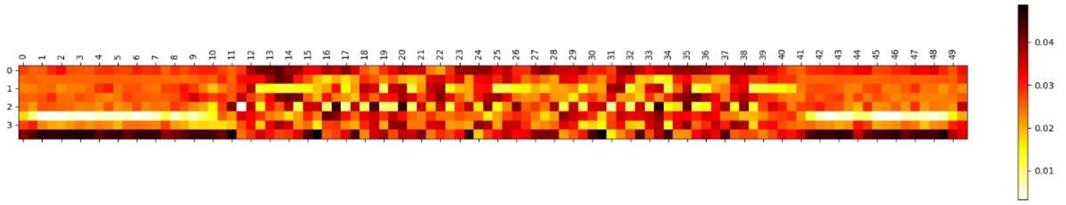


Figure 3. Attention visualization for an instance from sichuan dialect.

As shown, we can observe the attention is useful to conduct Chinese dialects discrimination. For Figure 2, this audio file is an example in shanghai dialect, and the number of frames is 280. After handling by our model, the number of frames is extracted into 70. The x-coordinate 0-34 represents the extracted 70 frames with interval 2, and the y-coordinate represents the 8 heads of the attention mechanism with interval 2. It can be seen that each head has a certain degree of access to audio information, and the discriminative phoneme of the frame number range from 20 to 56 have a great impact. Similarly, for Figure 3, this audio file is an example in sichuan dialect, the number of frames is 400. After handling by our model, the number of frames is extracted into 100 frames. The x-coordinate 0-49 represents 40 frames after extraction with interval 2, and the y-coordinate represents 8 heads of the attention mechanism with interval 2. It can be seen that the last one heads get the most information (with the deepest colour), and the features with frame number ranging from 26-76 have a great influence.

Although our model has achieved better performance in language discrimination tasks, the word error rate of our model in dialect speech recognition is still quite low. In the future, we will improve the performance of dialect speech recognition model through integrating more dialect corpus.

## 5. Conclusions

This paper designs an end-to-end dialect discrimination model based on multi-headed self-attention mechanism, which considers the influence of dialect pronunciation characteristics (phoneme sequence). In terms of dialect pronunciation features, we compared the pronunciation features extracted from different model structures. The experimental results on the benchmark speech corpus of 10 major dialects released by iFLYTEK demonstrate the effectiveness of the multi-headed self-attention mechanism, the performance of dialect discrimination has been greatly improved. We will further study how to better extract the unique pronunciation features of different dialects and design a composite model to further improve the performance of dialect discrimination. In the future, we will also expand our dialect corpus and focus on improving the performance of dialect speech recognition.

## ACKNOWLEDGEMENTS

The authors would like to thank anonymous reviewers for their insightful comments on this paper. This research was supported by the National Natural Science Foundation of China (NSFC) under Grant 61772246 and 61876074, Natural Science Foundation of Jiangxi Province under Grant 20192ACBL21030.

## REFERENCES

[1] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech communication, 2010, 52(1): 12-40.

[2] Li H, Ma B, Lee K A. Spoken language recognition: from fundamentals to practice[J]. Proceedings of the IEEE, 2013, 101(5): 1136-1159.

[3] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE signal processing letters, 2006, 13(5): 308-311.

[4] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[5] Lopez-Moreno I, Gonzalez-Dominguez J, Plchot O, et al. Automatic language identification using deep neural networks[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 5337-5341.

[6] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on audio, speech, and language processing, 2011, 20(1): 30-42.

[7] Richardson F, Reynolds D, Dehak N. A unified deep neural network for speaker and language recognition[J]. arXiv preprint arXiv:1504.00923, 2015.

[8] Ferrer L, Lei Y, McLaren M, et al. Study of senone-based deep neural network approaches for spoken language recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 24(1): 105-116.

[9] McLaren M, Ferrer L, Lawson A. Exploring the role of phonetic bottleneck features for speaker and language recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5575-5579.

[10] Mounika K V, Achanta S, Lakshmi H R, et al. An Investigation of Deep Neural Network Architectures for Language Recognition in Indian Languages[C]//INTERSPEECH. 2016: 2930-2933.

[11] Lozano-Diez A, Zazo-Candil R, Gonzalez-Dominguez J, et al. An end-to-end approach to language identification in short utterances using convolutional neural networks[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[12] Ma J, Song Y, McLoughlin I V, et al. LID-senone extraction via deep neural networks for end-to-end language identification[J]. 2016.

[13] Jin M, Song Y, McLoughlin I V, et al. End-to-end language identification using high-order utterance representation with bilinear pooling[J]. 2017.

[14] Cai W, Cai Z, Zhang X, et al. A novel learnable dictionary encoding layer for end-to-end language identification[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5189-5193.

[15] Garcia-Romero D, McCree A. Stacked Long-Term TDNN for Spoken Language Recognition[C]//INTERSPEECH. 2016: 3226-3230.

[16] Tkachenko M, Yamshinin A, Lyubimov N, et al. Language identification using time delay neural network d-vector on short utterances[C]//International Conference on Speech and Computer. Springer, Cham, 2016: 443-449.

[17] Pešán J, Burget L, Černocký J. Sequence summarizing neural networks for spoken language recognition[J]. Interspeech 2016, 2016: 3285-3288.

[18] Gonzalez-Dominguez J, Lopez-Moreno I, Sak H, et al. Automatic language identification using long short-term memory recurrent neural networks[C]//Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[19] Geng W, Wang W, Zhao Y, et al. End-to-end language identification using attention-based recurrent neural networks[J]. 2016.

[20] Gelly G, Gauvain J L. Spoken Language Identification Using LSTM-Based Angular Proximity[C]//INTERSPEECH. 2017: 2566-2570.

[21] Masumura R, Asami T, Masataki H, et al. Parallel phonetically aware DNNs and LSTM-RNNs for frame-by-frame discriminative modeling of spoken language identification[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 5260-5264.

[22] Tang Z, Wang D, Chen Y, et al. Phonetic temporal neural model for language identification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 26(1): 134-144.

[23] Gelly G, Gauvain J L, Le V B, et al. A Divide-and-Conquer Approach for Language Identification Based on Recurrent Neural Networks[C]//INTERSPEECH. 2016: 3231-3235.

[24] Fernando S, Sethu V, Ambikairajah E, et al. Bidirectional Modelling for Short Duration Language Identification[C]//INTERSPEECH. 2017: 2809-2813.

[25] Tang Z, Wang D, Chen Y, et al. Phonetic temporal neural model for language identification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 26(1): 134-144.

[26] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013: 6645-6649.

[27] Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015: 167-174.

[28] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 4835-4839.

[29] Yi J, Tao J, Bai Y. Language-invariant Bottleneck Features from Adversarial End-to-end Acoustic Models for Low Resource Speech Recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6071-6075.

[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[31] Zhou S, Dong L, Xu S, et al. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese[J]. arXiv preprint arXiv:1804.10752, 2018.

[32] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[33] Cai W, Cai Z, Zhang X, et al. A novel learnable dictionary encoding layer for end-to-end language identification[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5189-5193.

[34] Cai W, Chen J, Li M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[J]. arXiv preprint arXiv:1804.05160, 2018.

[35] "The 2015 NIST language recognition evaluation plan (LRE15)," NIST, 2015, ver. 22-3.

[36] Cai W, Cai Z, Liu W, et al. Insights into end-to-end learning scheme for language identification[J]. arXiv preprint arXiv:1804.00381, 2018.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# ONLINE ASSESSMENT OF ENGLISH FOR SPECIFIC PURPOSES

Renáta Nagy[1, 2] and Dr. Vilmos Warta[1]

[1]Department of Languages for Biomedical Purposes and Communication
Medical School
[2]Doctoral School of Health Sciences University of Pécs, Hungary

## ABSTRACT

*The study is about the questions of online assessment of English for Specific Purposes. The focus is on online assessment as a possible future form of language testing which truly has a huge importance hence today's situation around the globe. It can unquestionably be used as a perspective in a vast array of contents The study is innovative and its main aim is to uncover the intriguing questions and issues of online testing and to enlighten the candidates and professional assessors about the advantages and disadvantages of online testing. A thorough experimental process is currently being implemented involving a process that includes an online questionnaire completed by English and Hungarian students at the Medical School of the University of Pécs. Material and methods include already completed surveys, which will be followed by needs analysis and trial versions of online tests. These surveys do not only question future candidates but also assessors in order to find both perspectives of needs and wants. These include the aspects of tasks, content, skills, technology and others.*

## KEYWORDS

*Assessment, ESP, Language testing, Online, Validity.*

## 1. INTRODUCTION

Language testing is among the most interesting topics in the field of applied linguistics and it is also popular with foreign language teachers. Assessing language learners not only gives teachers feedback and information about the learners' knowledge but also gives a picture to the learners themselves regarding how well they have acquired the language and how well their learning skills work. Test results can also help as a motivating factor for sustaining and increasing students' willingness to learn (Turek, 1998). A considerable number of tests and examinations exist already regarding language acquisition, yet all fall victim to little variation. It has long been questioned whether it is of considerable use to vary *online* assessment when compared with traditional testing methods. Another intriguing question is whether it can be said that the evaluation of languages for specific purposes reflect today's current progressive attitudes, reflecting online behaviour. The number of questions arising is merely uncountable.

The present paper aims to provide an overview of the most up-to-date questions of online testing. The questions include the students' attitude towards online tests, their experience regarding the online world and their online habits. First, the paper will discuss the theoretical background providing relevant literature review of ESP and English in its general use, indicating the similarities and differences between them. Then, it will focus on assessment, more precisely online assessment. The literature of online testing has grown enormously in the past few months

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 317-325, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111426

hence the global pandemic situation. In this section it will also state the advantages and disadvantages of online assessment versus the traditional way of testing. Finally, it will draw conclusion of the survey that was carried out among the Hungarian and international students at the Medical School, University of Pécs.

## 1.1. Validity in the Online World

Validity is one of the key questions of online testing beside implementation. There have been numerous studies to investigate the intriguing field of validity. Czéreová and Mazurová (2017) claim that a test is considered valid if it can be used to measure what is supposed to measure. In the case of ESP, a test is considered valid if it measures not only the general language ability of the candidate, but it also measures the candidate's skills in certain field-specified context. Bachmann (2013) explains the term valid as a major requirement in order to be able to conduct a language assessment with a reliable result of the test taker's language ability.

Further investigations and experiments need to be conducted for an online test to be considered valid. Finding a way to prove validity of online testing is the next level of this study. In order to gain relevant information and solution about validity is our following concern.

## 1.2. Theoretical Background

The aim of the present paper is to uncover some of the most intriguing questions of online assessment of ESP. The topic has been on its peak since the pandemic situation made teachers and candidates stay at home behind their screens. Today's situation has forced people to look forward hence today's situation's urge to be able to fulfil everything online due to factors like time, progression and safety. There are so many ways of keeping up with the pace of technology as well as information technology itself, that somehow, we might as well miss the point and tend to think that everything is available online already. However, it remains a question whether we are ready for this challenge in terms of ESP.

At first, the term *Languages for Specific Purposes* should be clarified. It is still argued today whether Languages for Specific Purposes means merely having a certain set of vocabulary or simply having good communication skills result in being able to use a language well for any kind of purposes. According to Kurtán (2003), LSP includes elements of everyday language, common set of a specified vocabulary and defined specific vocabulary of a given field. Taking general and specified vocabulary into consideration, questions of assessment arise. It is questionable whether a language should be assessed in terms of vocabulary and if so, how we can differentiate between using a language, and more precisely a set of vocabulary for *general* purposes and using it for *specific* purposes. These are only the first questions that come to our mind when thinking about assessing ESP.

Numerous researchers have defined LSP so far. Kurtán (2003) provides a collection of the various descriptions of languages for specific purposes in linguistics. She describes it as a way of communication with the aid of verbal and non-verbal features, and she also claims that it can be used to communicate different messages from one to another participant and lastly, she states that we can consider it in a very narrow but also a very broad content.

Starfield (2013) states that in the beginning of teaching ESP, the motivation behind it was the need to communicate across the languages in areas such as commerce and technology. It can clearly be said that the initial motivation has not changed but numerous other and new topics have been added to the above mentioned areas, such as the areas of English for academic, social and business purposes.

If we take the functions of ESP into consideration, it is not difficult to agree with Basturkmen's (2006) claim about ESP, when she states that ESP has the following typical functions: help language learners to cope with the features of language or to develop the competences which are necessary to function in a discipline, profession or workplace. These functions are limited to a certain area, unlike when using a language in its general form. However, Hutchinson and Waters (1987) share their view of ESP as an *approach* rather than a *product*, by which they mean that ESP implicate a specific kind of language, teaching material or methodology. This view brings us back to the previous views regarding the close relation between ESP, profession and commerce.

As Master indicates about English and its role, it has a "subtle aspect of linguistic dominance" (1998:720) in ESP. It is unquestionable as English is still one of the most dominant languages around the globe, not only in spoken communication but in terms of written media and languages of specific purposes as well.

One of the main questions of the differentiation is whether or not we should make a distinction between Languages for Specific Purposes and languages in their 'general' use. Where is the line between the two above mentioned uses? If there is a line, is the line distinct or only a faint one. It is arguable, whether it is only the *function* as purpose that derives them form each other or they are separate due to the use of a special set of vocabulary. Warta (2005:28) explains this as follows: "The specialized language is different from the so-called general language regarding lexical, semantic, grammatical, stylistic, textual, sociolinguistic and pragmatic attitudes. Special vocabulary is only part of the language and linguistic repertoire needed for achieving special communication purposes". In this view, there are numerous features that draw a distinct line between language in its general and specific use. These features do not only appear in their everyday meaning, but also in this special, multi-purpose meaning.

As it is important in general EFL contexts, assessment is an integral part of LSP, too, which may be achieved through traditional as well as alternative ways (Bánhegyi-Fajt, 2020; Bánhegyi-Fajt-Dósa, 2020). Douglas (2000) claims that a distinction between EFL and LSP tests should be made. The fundamental features include authenticity, interaction and the knowledge of specific purpose content. In his view, for a task to be authentic, it is essential for the task to share critical features in the target language when measuring social skills in a situation exercise. The situation is in a specific context involving the criteria of the certain field. In this case, the language user is more likely to use the target language in a test situation as he would use it in a real-life situation. To be able to perform well in such exercise, good background knowledge is essential in the concept of specific purpose language use. In a case of a given field's situation, a specific set of vocabulary is needed to be able to complete such task on an expected level.

## 2. MATERIAL AND METHODS

As a first step, a query was conducted with the aid of a questionnaire on an online platform. There were two reasons behind choosing an online platform. Firstly, the topic itself is about an online issue, secondly, it is easier, quicker and more current to use an online platform with today's university students. The number of students answering the questionnaire made it clear that it is the best way of reaching out to students in such cases and topics.

The questionnaire listed twenty-one questions including questions about personal biodata, past online habits, present online habits, preferences regarding paper-based and online tests and attitude towards online tests and the attitude towards unethical behaviour.

This questionnaire is a first step in order to gain a relevant picture of online testing with the aim of a preliminary investigation to have a background picture from the students' point of view.

Their aspect is essential regarding further investigations, such as wants analysis. Eliciting students' attitude, needs and wants, experience and feelings about online assessment proves to be fruitful for this research.

In order to answer the research questions the quantitative research paradigm was used. The questionnaire was created using Google Forms, the answers were recorded and data was saved and personal data was kept confidential. Charts and MS Excel tables were used for data analysis, extraction and visualisation.

The second step is to investigate the assessors' attitude and expectations. To explore this side of the topic, a questionnaire will shortly be completed by me, which will be asked to complete by the ESP tutors at the Medical School, University of Pécs and by the Profex assessors. The aim of this is to get insight to the assessors' point of view, to receive a needs analysis in connection with assessing online. This will give another perspective on which path the research will gain valuable information to develop the experimental side. Online trial tests will give us the real, first-hand experience of the topic. It can only follow the thorough processing of the data received in the theoretical part. Volunteer participants will again come from the same basis, volunteering students of the University of Pécs Medical School.

A great deal of organization needs to be done in advance, which include finding a suitable room with reliably working computers, an applicable software, camera system, personnel, eligible tests, given personality rights for video recording and the list goes on. Once the online trial test is prepared thoroughly and commenced successfully, it will give a huge set of information urging to be processed to find out about the advantages and disadvantages of online testing.

### 2.1. Participants

Participants included Hungarian and international students at the Medical School, University of Pécs. In this particular questionnaire, the participation was limited to students of higher education and to the students of the University of Pécs. The reason behind this is to acquire answers and data from one particular educational age group. To find out about different age groups' attitude and answers further investigations are necessary with the possibility of comparing the given data.

## 3. DISCUSSION

Data, regarding both students in higher education and ESP can be seen on the following charts. (Figure 1.) In our case, Medical English is behind the meaning of specific purpose. At the Medical School of the University of Pécs, candidates have the possibility to take Profex language exam, which is an exam for languages in Medical Purposes. "*PROFEX (PROficiency EXamination) in an English for Medical Purposes (EMP) bilingual testing system offering language tests for medical and paramedical professionals.*"(http://profex.aok.pte.hu/en/what-profex) There are three levels B1, B2 and C1, which were defined in accordance with the recommendations of the Council of Europe described in the Common European Framework of Reference for Languages. The PROFEX Exam Center is at the Faculty of General Medicine, University of Pécs. PROFEX exams are administered in more than 20 exam sites in Hungary and abroad too, for example in Targu Mures University of Medicine and Pharmacy in Romania. Exams can be taken in four languages and successful exams are certified by official language certificates.

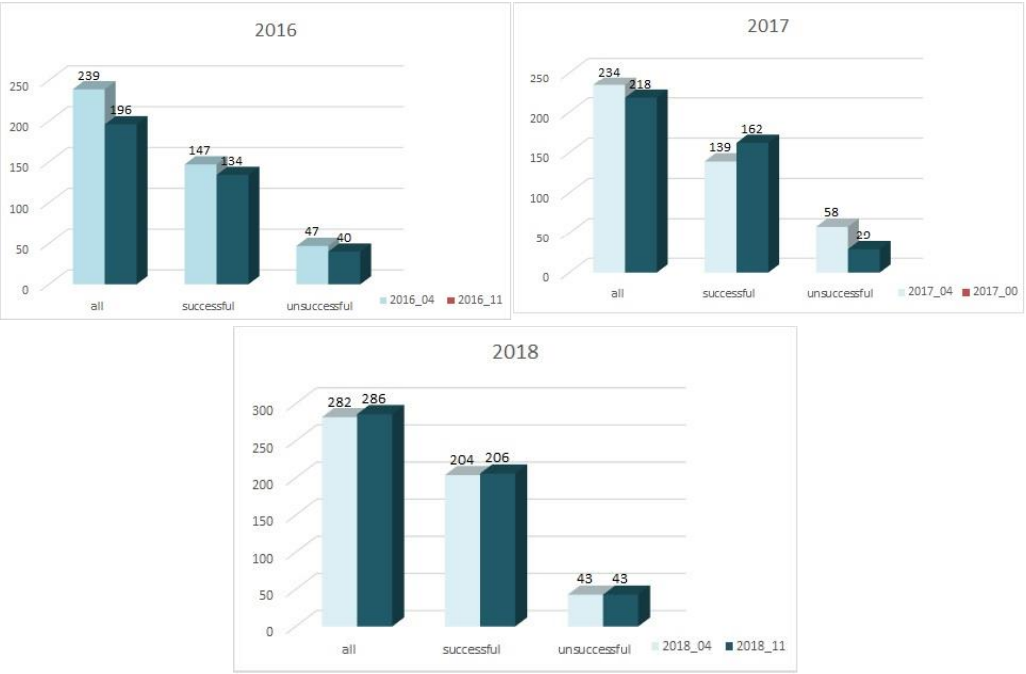


Figure 1. Number of successful and unsuccessful Profex candidates 2016-2018. English (lighter bar) and German (darker bar) for Specific Purposes (http://profex.aok.pte.hu/en/what-profex)

The above charts show the number of candidates taking language exams for Medical Purposes. It is clearly visible that there is a increasing demand for these exams year by year. The total number of candidates of Profex language exams show a steadily rising tendency in the last three years' statistics. It is also noticeable that the number of unsuccessful candidates is significantly fewer compared to the successful ones. From 2016 to 2018 the total number of candidates who applied increased from 435 to 586. The number of successful candidates also shows a positive dynamic with the rise from 281 to 410. As for the unsuccessful candidates, we can see a merely steady number of 87 and 86 students. This exam has worked offline so far, thus it is an interesting question how these numbers would change if there was an opportunity to take it online. One of the most important questions and obstacle of this at the same time is validity.

## 4. Results

Having identified students' experience, attitude and view on online assessment has given us insightful findings, however it also showed that further examination is required to handle today's increasing online realm.

The questionnaire's twenty-one questions were answered by 430(n=430) students of the Medical School at the University of Pécs. The students included Hungarian and international students. According to the nationality figure, 31 nationalities participated in the investigation. Figures 2 and 3 gives us insight to the students' answers to the 2 basic questions of the survey. A large pool of participants has been involved in an online test before, as 74.4% (n=320) of the students chose yes as an answer to that question. The answer to the follow-up question also received yes from the vast majority of the students as 84.9% (n=365) clicked on that option and only 25.1% (n=65) clicked on no. These results indicate that students already have experience with online tests to some extent and it also indicates students' willingness to try to gain more of it.

Interestingly, after these two high percentage figures of online tests, another follow-up question only received 55.8% (n=240) yes answers for the following question: Do you believe online tests should be available on a broader scale in consideration of the 21st century?

The last figure of answers shows us that even though the present generation is open and has experience with the online world yet fall victim to the unknown world of online assessment. They clearly have more practice in the traditional, paper-based tests which gives them confidence in that type of assessing method. To our knowledge, this is the first report on the online and the traditional ways of testing which gives us ground for a future research. In that research the potential effects of having much more online exams due to today's situation should be taken into consideration more carefully.

https://docs.google.com/forms/d/1CoyxMyYXjnD2r8vwOc_OIi6RpQVO46GKsmoXiYODVIA/edit?no_redirect&gxids=7628#responses

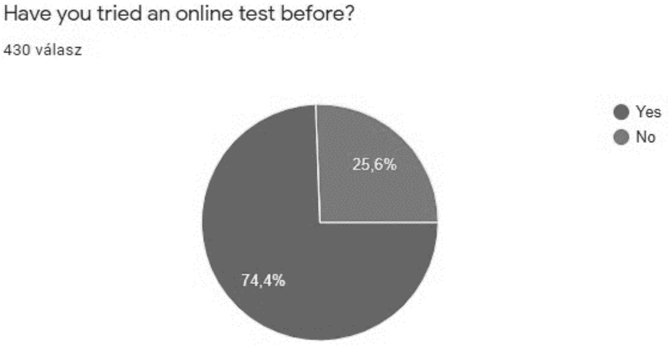


Figure 2. Percent of students who has tried an online test before

https://docs.google.com/forms/d/1CoyxMyYXjnD2r8vwOc_OIi6RpQVO46GKsmoXiYODVIA/edit?no_redirect&gxids=7628#responses

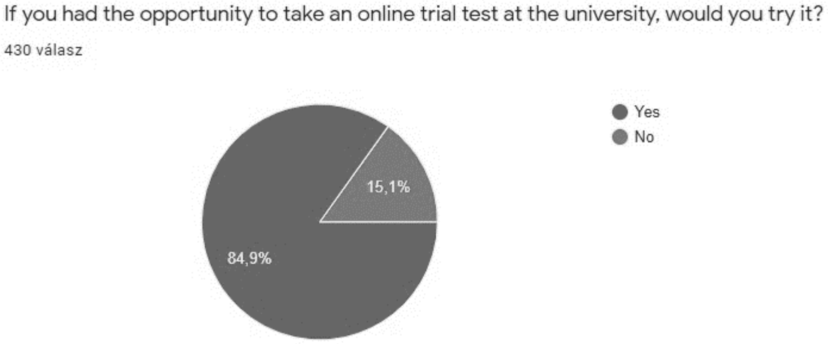


Figure 3. Percent of students who would try an online trial test at the university if there was an opportunity

In the research, one of the questions invesigated the fields of life in which students would prefer online tests to traditional tests if they had the opportunity for the online version. The following diagram (Figure 4) shows us the first four items receiving the top votes. They are the following: the theoretical part of the driving course with 64.9% (n=279), computer skills with 52.1% (=224), academic staff with 50.5% (=217) and language certificate with 49.8% (=214). These findings indicate that the students prioritise those four fields and that they would prefer to have these tests available online. The broad implication of the present result to this question is that in the previously mentioned fields are of high importance for this generation, thus their online test versions are worth researching in the future.

https://docs.google.com/forms/d/1CoyxMyYXjnD2r8vwOc_OIi6RpQVO46GKsmoXiYODVIA/edit?no_redirect&gxids=7628#responses

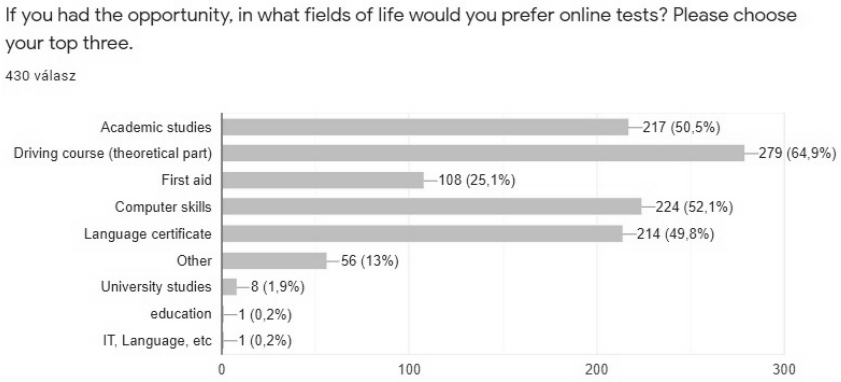


Figure 4: Fields in which students would prefer online tests to traditional tests

The last graph of this paper (Figure 5) shows us the students' top three chosen characteristics regarding online tests. The most popular feature of online tests among the students was "it can be done anywhere" which received a significant 84.2% (n=362). This suggests that being able to take a test at any place is among the most important feature. This was followed by the answer "it can be done anytime" which received 75.8% (n=326). This implicates that the time factor is almost as important for possible candidates as the place factor. Another time-related answer received the third most significant percentage, namely "I would get the results sooner" 74.2% (n=319). These interesting research questions indicate that time and place are of high importance in for today's generation and it also gives us opportunity for future research that can be derived from the findings of the present study.

https://docs.google.com/forms/d/1CoyxMyYXjnD2r8vwOc_OIi6RpQVO46GKsmoXiYODVIA/edit?no_redirect&gxids=7628#responses

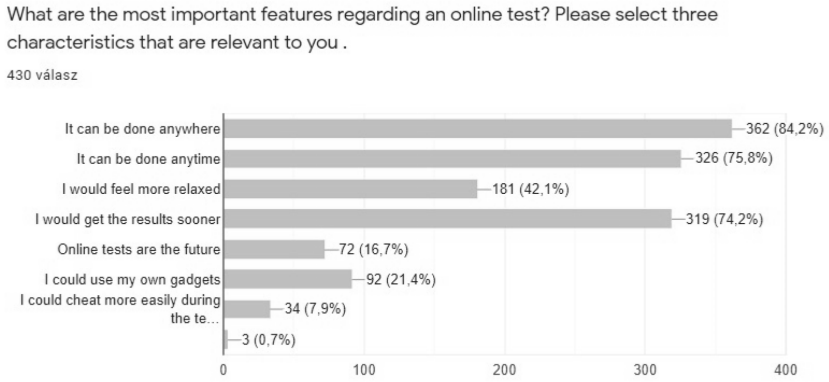


Figure 5: Most important features of online tests for students

## 5. CONCLUSION

One of the characteristic goals in support of this study is to outline a comprehensive vision of ESP assessment from varying perspectives, including context, time, place, tasks, test developing and technology. However, we acknowledge that there are considerable researches should still be conducted to gain a relevant picture in this field after facing several months of online experience.

I intend to do this not only from the aspect of theory but also from first-hand experience. Throughout the entirety of this research, my primary goal is to identify such perspectives of online testing. I intend to achieve results which can be an aid for actively assessing teachers and also for students desiring to improve their ESP skills aiming at an improving level to conduct a reliable and valid online assessment. The present findings confirm the need and value of online assessment, however the concern of validity and technical issues provide a good starting point for discussion and further research for us to be able to move on to the next level.

### ACKNOWLEDGEMENTS

I am most grateful for the professional aid of my consultant, Dr. Vilmos Warta, for the artistic help of Jon E. Marquette and I am grateful for the patience of my three wonderful children.

### REFERENCES

[1] Bachmann, L.F. (2013): Ongoing Challenges in Language Assessment. The Companion to Language Assessment, First Edition. Ed: Antony John Kunnan. John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla128

[2] Bánhegyi M. – Fajt B. (2020): Portfólió a gazdasági szaknyelvoktatásban: hallgatói attitűdök. Modern Nyelvoktatás. 26/3. 38-50

[3] Bánhegyi, M. - Fajt, B. - Dósa, I. (2020): Szaknyelvi portfólió újratöltve: egy hallgatói attitűdfelmérés tapasztalatai. In: Bocz, Zs. – Besznyák, R. (eds) (2020): Porta Lingua 2020. Szaknyelvoktatás és -kutatás nemzetközi kontextusban. Cikkek, tanulmányok a hazai szaknyelvoktatásról. SZOKOE: Budapest. 215-230. DOI: https://doi.org/10.48040/PL.2020.18

[4] Basturkmen, H (2006): Ideas and options in English for specific purposes. Lawrence Erlbaum Associates: New Jersey

[5] Czéreová, B. – Mazurová, H. (2017): Assessment as an Integral Part of Teaching Business English. Porta Lingua: Budapest

[6] Douglas, D. (2000): Assessing Languages for Specific Purposes. Cambridge University Press: Cambridge

[7] Hutchinson, T. – Waters, A. (1987): English for Specific Purposes. Cambridge University Press: Cambridge

[8] Kurtán, Zs. (2003): Szakmai Nyelvhasználat. Nemzeti Tankönyvkiadó: Budapest

[9] Master, P. (1998): Positive and negative aspects of the dominance of English. TESOL Quarterly 32. 716-27.

[10] Paltridge, B. – Starfield S. (eds) (2013): The Handbook of English for Specific Purposes. Wiley-Blackwell: Boston

[11] Warta, V. (2005): Szerzői hang: Angol nyelvű orvosi esetismertetések műfajelemzése korpusznyelvészeti módszerekkel. PhD értekezés. Kézirat. Pécsi Tudományegyetem: Pécs Internet site references

[12] Profex – State Recognised Language Examination Center. http://profex.aok.pte.hu/en/what-profex. [Access: 2021.01.31.]

## AUTHOR

My name is **Renáta Nagy** and I am an assistant lecturer at the Department of Languages for Biomedical Purposes and Communication at the University of Pécs Medical School, Hungary. I am an English and Hungarian for Specific Purposes instructor and my primary areas regarding classroom instruction include General English, Medical English, Medical Hungarian and Terminology. Recently, I initiated my PhD at the Doctoral School of Health Sciences (University of Pécs) in which my field of research includes the question of online assessment.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# LOW-RESOURCE NAMED ENTITY RECOGNITION WITHOUT HUMAN ANNOTATION

Zhenshan Bao, Yuezhang Wang and Wenbo Zhang

College of Computer Science, Beijing University of Technology, Beijing, China

## ABSTRACT

*Most existing approaches to named entity recognition (NER) rely on a large amount of high-quality annotations or a more complete specific entity lists. However, in practice, it is very expensive to obtain manually annotated data, and the list of entities that can be used is often not comprehensive. Using the entity list to automatically annotate data is a common annotation method, but the automatically annotated data is usually not perfect under low-resource conditions, including incomplete annotation data or non-annotated data. In this paper, we propose a NER system for complex data processing, which could use an entity list containing only a few entities to obtain incomplete annotation data, and train the NER model without human annotation. Our system extracts semantic features from a small number of samples by introducing a pre-trained language model. Based on the incomplete annotations model, we relabel the data using a cross-iteration approach. We use the data filtering method to filter the training data used in the iteration process, and re-annotate the incomplete data through multiple iterations to obtain high-quality data. Each iteration will do corresponding grouping and processing according to different types of annotations, which can improve the model performance faster and reduce the number of iterations. The experimental results demonstrate that our proposed system can effectively perform low-resource NER tasks without human annotation.*

## KEYWORDS

*Named entity recognition, Low resource natural language processing, Complex annotated data, Cross-iteration.*

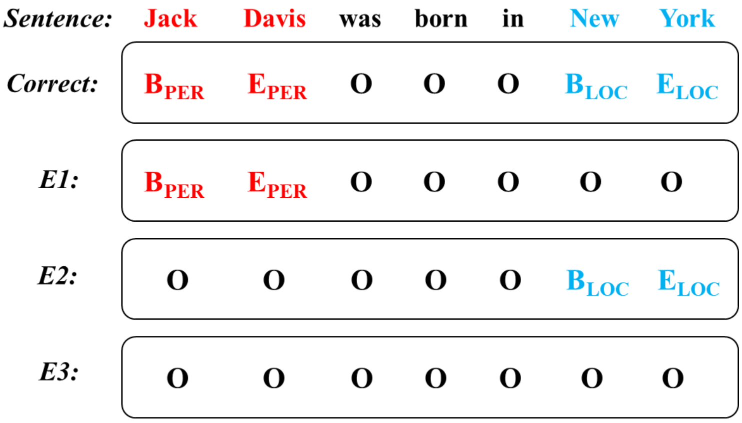
## 1. INTRODUCTION

Named entity recognition is widely applied in many scenarios, usually as a basic task for information extraction, question answering and machine translation [1]. Most existing approaches to NER focused on a supervised setup, which rely on a large amount of high-quality annotations [2]. However, in practice, it is very expensive to obtain manually annotated data. In most cases, a list of entities will be constructed to annotate the data automatically, which has high requirements for the comprehensiveness of the entity list. In some special fields, it is difficult to provide a more comprehensive list of entities. In this low-resource situation, the data automatically annotated using the entity list is often not perfect, including incomplete annotation data or non-annotated data.

Figure 1 shows an example of annotating data with a list of entities. The sentence with two named entities “Jack Davis” and “New York” of type PER (person) and LOC (location), respectively. Following the standard “BIOE” annotation system, the correct annotations is shown

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 327-335, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111427

below the sentence. In a real scenario, the provided data annotations may be missing or incorrect, especially automatic annotations. Examples 1 and 2 (E1 and E2) are incomplete data annotations, not all entities in the corpus are annotated. The corpus with no entity annotated is non-annotated data, as shown in E3. The entity of non-annotated data may not be annotated or there may be no entity in the corpus.



| *Sentence:* | Jack | Davis | was | born | in | New | York |
|---|---|---|---|---|---|---|---|
| *Correct:* | $B_{PER}$ | $E_{PER}$ | O | O | O | $B_{LOC}$ | $E_{LOC}$ |
| *E1:* | $B_{PER}$ | $E_{PER}$ | O | O | O | O | O |
| *E2:* | O | O | O | O | O | $B_{LOC}$ | $E_{LOC}$ |
| *E3:* | O | O | O | O | O | O | O |

Figure 1. Example of using entity list to annotate data

It is difficult to obtain high performance when these noisy data are directly used in the training model. For these annotations, previous work mainly focused on one of these annotation data types. Using these complex data to train the NER model is a very challenging task. Solving this problem can effectively reduce the application difficulty of NER tasks in actual scenarios and reduce costs.

In this work, we present a flexible and efficient system to deal with complex data automatically annotated by entity list, and effectively use these data to improve the performance of NER model. Our system extracts semantic features from a small number of samples by introducing a pre-trained language model. Based on the incomplete annotations model [16], we relabel the data using a cross-iteration approach. We use the data filtering method to filter the training data used in the iteration process, and re-annotate the incomplete data through multiple iterations to obtain high-quality data. Finally, we use these re-annotated data to train the final NER task model. To evaluate the efficiency of our system, we conduct experiments on two real network datasets. The experimental results demonstrate that our proposed system can effectively perform low-resource NER tasks without human annotation.

## 2. Related Work

Traditional NER methods are mainly rule-based methods and statistical-based methods. Alfred et al. and Hanisch et al. [3, 4] perform NER tasks through rule matching or grammar rules. Statistics-based methods such as Hidden Markov Model (HMM) [5], Max Entropy Model [6], Conditional Random Fields (CRF) [7] and Support Vector Machine (SVM) [8] are also classic methods for processing NER tasks. In recent years, deep learning has developed rapidly, and neural networks are usually used for NER tasks. Chiu et al. and Zhang et al. [9, 10] use neural network model to obtain character-level or word-level representations from large amounts of annotated data.

Although the named entity recognition method based on deep learning has achieved good results, most of the current deep learning models with good recognition performance often rely on a large number of high-quality annotated data. To solve this problem, Peters et al. [11] use the pre-trained context embedding of the language model to perform the NER task through a semi-supervised method. BERT [12] shows great advantages and achieve substantial performance improvements, and pre-trained language models have become a very important component in recent. Helwe et al. [13] proposed a co-training approach, while adding Arabic-based word embedding, using a small amount of data to improve the performance of the model.

The processing of noise annotations in NER task has attracted attention. Lou et al. [14] proposed a dictionary-based graph attention model to deal with this problem. Yang et al. [15] presents an approach to utilize the data generated by distant supervision to perform newtype named entity recognition in new domains. Jie et al. [16] use re-annotation method to solve this problem, the annotations in the next iteration are re-annotated by the model learned in the previous iteration. Peng et al. [17] propose an algorithm that uses only un-annotated data and a list of named entities to process NER tasks, the PU algorithm. In order to process tokens with a variety of possible tags, AutoNER [18] proposed an improved fuzzy CRF layer for processing to improve the performance of the model.

## 3. APPROACH

This work is aimed to improve the performance of NER systems by inferring missing information from a small number of entity lists. We propose an efficient and flexible system to deal with complex data automatically annotated by entity list, and effectively use these data to improve the performance of NER model. Our NER system mainly includes three modules, data annotation, NER model, and cross-iteration approach using data filtering methods.

Our system does not use any human-annotated data for model training, so we use a small list of entities to automatically annotate the training data. We use the BERT-CRF model as the basic model of our NER system. Inspired by Jie et al. [16], we propose a cross-iteration approach and data filtering method to re-annotate imperfect training data. Finally, we use these re-annotated high-quality training data to train the final model to improve the performance of the NER system.

### 3.1. Data Annotations

For NER tasks, the most common knowledge is a list of entities describing many entities belonging to the same category. Entity lists are relatively cheap, because there are many existing lists, and if coverage requirements are not high, it is easy to create an entity list manually.

The purpose of our system is to effectively utilize the automatically annotated data under low resources, so our entity list contains only a few entities. The entity list we used only included about 30% of the entities in the unlabeled corpus. We use the obtained entity list, use the forward maximum matching algorithm, and follow the standard BIOES tagging scheme to automatically annotate the entities in the sentence. After completing the annotation, words that are not entities and entity words that are not included in the entity list are both annotated as ‘O’.

In addition, we add an extra label to each word to record whether the word is annotated as an entity by the entity list during the data annotation process. This additional label will be used in the subsequent cross-iteration approach and re-annotation process. We believe that the entities annotated with the entity list are more accurate than the iterative model predictions, so recording these words can improve the efficiency and accuracy of the cross-iteration process.

## 3.2. Model Architecture

The model architecture we use is a classic NER model, using the BERT [12] model to obtain word embeddings and extract features, and then use a CRF layer to obtain the output tag sequence, as shown in Figure 2.

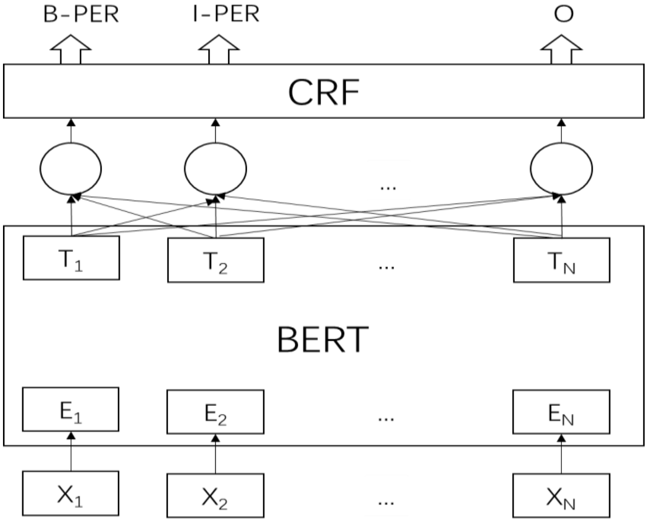


Figure 2. Model Architecture

In the task of NER, the pre-trained language model can be used to obtain the embedded representation of each word in the sentence, which can more accurately represent the semantic relationship between the entity and the context. BERT [12] is a model proposed by Google, which adopts Transformer [19] and self-attention mechanisms to learn contextual relations between words in a text, and has achieved outstanding results for a lot of NLP tasks. We use the BERT pre-training model to encode the word sequence as word embedding, and effectively mine the potential semantic information between the entity and the context, while reducing the number of samples required for training.

On top of the BERT model, a sequential CRF layer is used to perform label inference. The CRF layer can effectively constrain the dependency relationship between the predicted tags, thereby obtaining the global optimal sequence.

## 3.3. Cross-Iteration

Most of the annotations obtained by automatic annotation using entity list are imperfect, and direct use of these low-accuracy annotations will affect the performance of the model. Therefore, this paper proposes a cross-iteration approach and data filtering method to improve the entity annotation coverage of training data. Algorithm 1 shows the cross-iteration training procedure to re-annotate the data.

Firstly, we initialize the annotation data, record the annotation marked by entity list, and divide the training data into two folds. Then, we perform cross-iterations on the two training sets. Specifically, each time we train the model with half of the training data to predict the entity distribution of the other half of the training data. We re-annotated the imperfect annotations based on the predicted results. In order to improve the accuracy of re-annotated entities, we only re-annotate entities that were originally labeled as 'O'. At the same time, we reset the parameters of the model before each iteration to avoid accumulating more errors in the re-annotation process. After several iterations, the coverage of entities in the training data is improved and the entity information is enriched. Finally, use these re-annotated training data to train the final model.

In the iterative process, we find that the non-annotated training data has little contribution to the prediction. Therefore, we used the data filtering method to regroup and filter the data used to train the prediction model. In the data initialization phase, we group the training data according to whether the data is non-annotated data. On this basis, each group is equally divided into two folds for cross-iteration. In addition, filter out the non-annotated data in this fold before each iteration for the training of the prediction model.

When training the final model with the re-annotated data, we do not use the data filtering method. This is because the data has been more accurately annotated during the cross-iteration process, so the non-annotated data can be used to train the model.

**Algorithm 1: Cross-Iteration Training Procedure.**

```
Input:
$N$: number of iterations; $D_i$: incompletely annotated data; $D_n$: non-annotated data; $M$: model
Output:
$D_r$: re-annotated training data
Initialize the annotation data, record the annotation marked by entity list;
Divide $D_i$ into two folds $D_{ia}$ and $D_{ib}$;
Divide $D_n$ into two folds $D_{na}$ and $D_{nb}$;
Save initial parameters of model $M$ as $M_{init}$;
for iteration=1, 2, 3 … to $N$ do
    Filter out the non-annotated data after relabeling in $D_{na}$, and merge with $D_{ia}$ to get training data $T_a$
    Filter out the non-annotated data after relabeling in $D_{nb}$, and merge with $D_{ib}$ to get training data $T_b$
    Reset the parameters of the model $M$ to $M_{init}$;
    Train model $M$ with $T_a$, get model $M_a$;
    Train model $M$ with $T_b$, get model $M_b$;
    Use model $M_a$ to predict the $D_{ib}$ and $D_{nb}$;
    Use model $M_b$ to predict the $D_{ia}$ and $D_{na}$;
    Re-annotate $D_i$ and $D_n$ according to the prediction results, get re-annotated training data $D_r$;
end for
```

# 4. EXPERIMENTATION

## 4.1. Dataset and Experimental Settings

To evaluate the efficiency of our system, we conduct experiments on two real network datasets, AutoIE [21] and E-commerce-NER [22]. The corpus of the AutoIE dataset is derived from the title text of Youku video, which contains 10,000 samples without annotations for training and 1000 samples with fully annotated for testing. Besides the corpus, three lists of interested entity types are provided. These entities may cover around 30% entities occurring in the unlabeled corpus. The E-commerce-NER data set is a dataset crawled through the web and manually annotated. It contains two types of entities, namely products and brands. We randomly selected 30% of the annotated entities as the entity list, deleted all the original entity annotations, and re-annotated the training data with the entity list. Table 1 shows the distribution of non-annotations and incomplete annotations after using the entity list annotation.

Table 1. Statistic of dataset

| Datasets | Non-annotations | Incomplete annotations | Test |
|---|---|---|---|
| AutoIE | 5667 | 4333 | 1000 |
| E-commerce-NER | 574 | 3415 | 498 |

We compare our approach with the following model in NER task with incomplete annotation:

- Origin. It is a model that directly uses the data annotated by entity list for training;
- O-Filter. Filter out non-annotated data and use it to train the model. This model is used to evaluate the effectiveness of data filtering method.
- Baseline Jie et al. [16]: The system achieves state of art result for incomplete annotations problem in NER application, and it is employed as the baseline system for our evaluation.

We use the BERT pre-trained model “chinese_wwm_ext” which released by Cui [20]. Adam optimizer is used with the learning rate of 1e-3, and set batch size as 128. The epoch of each iteration prediction model is set to 20, and the number of iterations is set to 30.

## 4.2. Results

The best F1 results achieved by different methods on different datasets are listed in Table 2.

Table 2. Performance comparison between different methods

| Datasets | Model | Precision | Recall | F1 score |
|---|---|---|---|---|
| AutoIE | Origin | 0.7670 | 0.2981 | 0.4293 |
| | O-Filter | 0.7435 | 0.5577 | 0.6373 |
| | Baseline | 0.6435 | 0.6680 | 0.6555 |
| | Ours | 0.7436 | 0.7902 | 0.7662 |
| E-commerce-NER | Origin | 0.6121 | 0.5006 | 0.5508 |
| | O-Filter | 0.6107 | 0.5785 | 0.5942 |
| | Baseline | 0.5916 | 0.8044 | 0.6818 |
| | Ours | 0.6068 | 0.8391 | 0.7043 |

As shown in Table 2, our model can achieve the best performance on both datasets. The Origin model, which directly uses the data annotated by entity list for training, has poor performance. It is mainly due to the noise effect of a large number of non-annotated data in the training data. Therefore, we filter out non-annotated data for training. From the results of the O-Filter model, we can see that the performance of the model has been improved. Compared with the baseline model, our proposed method has achieved higher performance. On the one hand, our method uses a cross-iteration approach, which effectively utilizes the existing entity list and the semantic features of the entities in the corpus, which enhances the coverage and diversity of the entities in the training data. On the other hand, our data filtering method is used in each iteration process, reducing the accumulation of noise and further improving the accuracy of the prediction results.

On the AutoIE dataset, the best F1 score obtained by our method can reach 0.7662, which is 0.1107 higher than the baseline model. Compared with the baseline model for incomplete annotations, our method significantly improves the precision and recall, reaching 0.7436 and 0.7902, respectively, an increase of 0.1001 and 0.1222.

On the E-commerce-NER dataset, the best F1 score can reach 0.7043, which is also higher than other methods, but the performance improvement is not obvious on the AutoIE dataset. This is because in the E-commerce-NER dataset, the proportion of non-annotated data is relatively small, and the improvement of model performance by data filtering methods is also reduced.

In order to further explore the influence of data filtering method on the iteration approach, we draw the performance curve of the test model when iterating on the AutoIE dataset, as shown in Figure 3. The test model is a model trained using all the re-annotated data after each iteration.

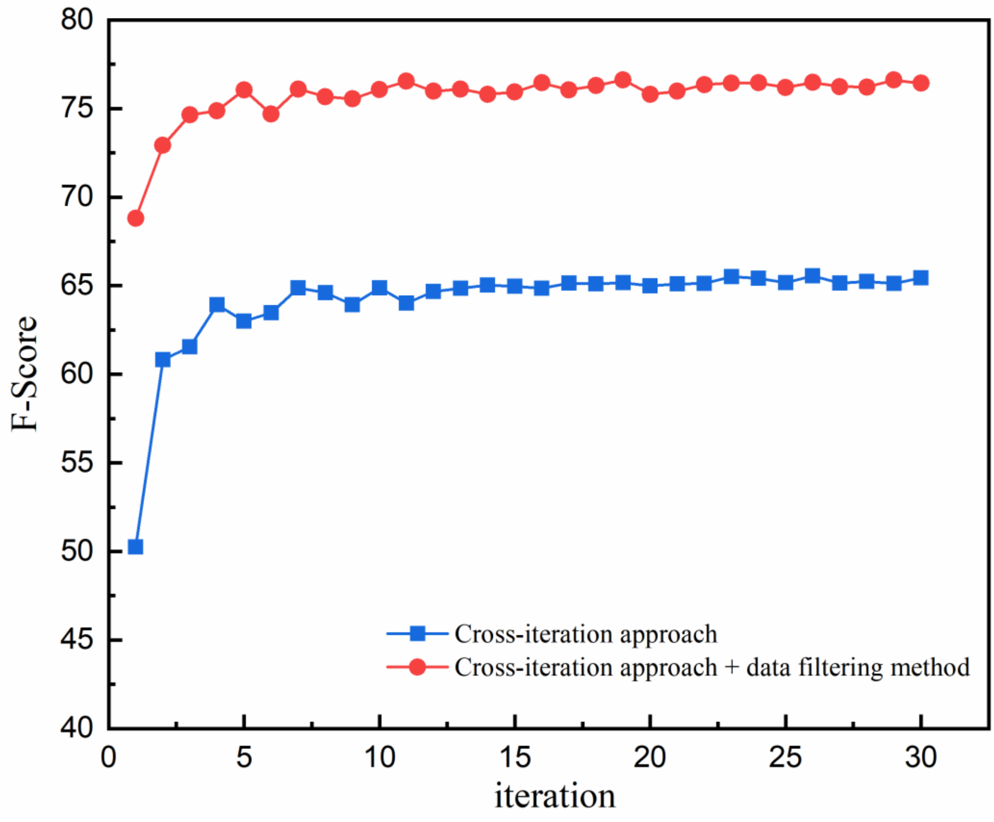


Figure 3. F1 score of iterative process on AutoIE dataset

We can see that the cross-iteration approach using the data filtering method can improve the model performance faster. For the training data automatically annotated with the entity list, the data filtering method can effectively filter out the noisy data, and at the same time re-add them to the training process through multiple cross-iterations to effectively use the data. By combining cross-iteration approach with data filtering method, we can get a higher performance model and greatly reduce the number of iterations required.

## 5. Conclusions

In this work, we designed an efficient and flexible system that uses automatic annotation data for NER tasks in a low-resource environment. We propose a cross-iteration approach and data filtering method to improve the entity annotation coverage of training data. Each iteration will do corresponding grouping and processing according to different types of annotations, which can improve the model performance faster and reduce the number of iterations. Experiments on real datasets show that our system significantly improves the performance of the NER model in the case of complex annotations. In future work, we will try to filter these complex training data in more detail, and we believe that our method can also be used for more routine tasks besides sequence annotation.

## ACKNOWLEDGEMENTS

This work is supported by National Key R&D Program of China (No. 2017YFC0803300), Beijing Natural Science Foundation under Grant 4172004, Beijing Municipal Education Commission Science and Technology Program under grant number KM201910005027.

## REFERENCES

[1] Yadav, V., & Bethard, S. (2018, August). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2145-2158).

[2] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering.

[3] Alfred, R., Leong, L. C., On, C. K., Anthony, P., Fun, T. S., Razali, M. N. B., & Hijazi, M. H. A. (2013, December). A rule-based named-entity recognition for malay articles. In International Conference on Advanced Data Mining and Applications (pp. 288-299). Springer, Berlin, Heidelberg.

[4] Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R., & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. BMC bioinformatics, 6(1), 1-9.

[5] Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC), 1(4), 15-23.

[6] Curran, J. R., & Clark, S. (2003). Language independent NER using a maximum entropy tagger. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 164-167).

[7] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[8] Ju, Z., Wang, J., & Zhu, F. (2011, May). Named entity recognition from biomedical text using SVM. In 2011 5th international conference on bioinformatics and biomedical engineering (pp. 1-4). IEEE.

[9] Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4, 357-370.

[10] Zhang, Y., & Yang, J. (2018, July). Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1554-1564).

[11] Peters, M., Ammar, W., Bhagavatula, C., & Power, R. (2017, July). Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1756-1765).

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[13] Helwe, C., & Elbassuoni, S. (2019). Arabic named entity recognition via deep co-learning. Artificial Intelligence Review, 52(1), 197-215.

[14] Lou, Y., Qian, T., Li, F., & Ji, D. (2020). A Graph Attention Model for Dictionary-Guided Named Entity Recognition. IEEE Access, 8, 71584-71592.

[15] Yang, Y., Chen, W., Li, Z., He, Z., & Zhang, M. (2018, August). Distantly supervised ner with partial annotation learning and reinforcement learning. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2159-2169).

[16] Jie, Z., Xie, P., Lu, W., Ding, R., & Li, L. (2019, June). Better modeling of incomplete annotations for named entity recognition. In Proceedings of the 2019 Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 729-734).

[17] Peng, M., Xing, X., Zhang, Q., Fu, J., & Huang, X. J. (2019, July). Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2409-2419).

[18] Shang, J., Liu, L., Gu, X., Ren, X., Ren, T., & Han, J. (2018). Learning Named Entity Tagger using Domain-Specific Dictionary. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2054-2064).

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[20] Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101.

[21] Yang, X., Wu, B., Jie, Z., & Liu, Y. (2020, October). Overview of the NLPCC 2020 Shared Task: AutoIE. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 558-566). Springer, Cham.

[22] Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., & Si, L. (2019, July). A neural multi-digraph model for Chinese NER with gazetteers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1462-1467).

## AUTHORS

**Bao Zhenshan** is an associate professor in faculty of information technology (FIT), Beijing University of Technology (BJUT), Beijing, China. His research interests include machine learning, natural language processing and their applications.

**Wang Yuezhang** received the Bachelor degree of Computer Science and Technology in BJUT in 2018. Now he is the postgraduate student majoring in computer Technology in BJUT. His current research is about deep learning, natural language processing, and low-resource named entity recognition.

**Zhang Wenbo** received her Ph.D. degree of Computer Science and technology in 2015. Now she is a lecturer in FIT, BJUT, Beijing, China. Her research interests include natural language processing, heterogeneous computing, intelligent computing system and their applications. And she is the corresponding author.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# AN INTELLIGENT MOBILE APPLICATION TO AUTOMATE THE ANALYSIS OF FOOD CALORIE USING ARTIFICIAL INTELLIGENCE AND DEEP LEARNING

Yongqing Yu[1], Yishan Zou[2] and Yu Sun[3]

[1]BASIS International School Park Lane Harbour, Huizhou, China 516001
[2]Department of Education, University of Pennsylvania,
Philadelphia, PA, USA, 19104
[3]Department of Computer Science, California State Polytechnic University,
Pomona, CA, USA 91768

### *ABSTRACT*

*As obesity becomes increasingly common worldwide [9], more and more people want to lose weight – for both their health and their image. According to the Centers for Disease Control and Prevention (CDC), long-term changes in daily eating habits (such as regarding food/nutrition type, calorie intake) are successful at keeping weights off [10]. Therefore, it would be helpful to have an AI mobile program that identifies the types of food the user consumes and automatically calculates the total calories. This paper examines the development and optimization of an 11-categorical food classification model based on the MobileNet neural network using Python. Specifically, it classifies any food image as one of bread, dairy, dessert, egg product, fried food, meat, noodles, rice, seafood, soup, or fruit/vegetables. Methods of optimization include data preprocessing and learning rate and batch size adjustments. Experimental results show that scaling image inputs to standard size (Python Numpy resize() function), 300 training epochs, dynamic learning rate (start with 0.001 and *0.1 for every 30 epochs), and a batch size of 16 yields our best model of 83.44% accuracy.*

### *KEYWORDS*

*Food classification, Python, Data preprocessing, MobileNet, Epochs, Overfitting, Learning rate, Batch size.*

## 1. INTRODUCTION

With the development of modern society, people's living standards are improving day by day, and the demand for dietary nutrition and health has also increased [11]. With the gradual popularization of scientific knowledge of dietary nutrition, dietary nutrition and health are gradually understood and pursued by the general public. On the one hand, various nutrients required by the human body can be obtained through diet, such as protein, fat, vitamins, water and inorganic salts, etc. On the other hand, the food intake by the human body should be suitable for digestion and absorption, and the food should be fresh. Pollution and pollution-free, food processing should be scientific and reasonable, nutrition should be ensured, appetite should be improved, taste and flavor should be ensured, and calorie intake of three meals a day should also be properly allocated. Based on the suggestions provided by the nutrition associations, the energy distribution at 3 meals per day accounts for 30%, 40%, and 30% of the whole day respectively.

DOI: 10.5121/csit.2021.111428

Most people with obesity lack nutritional knowledge, and the male are suffering a more serious issue than the female, so the problem of dietary malnutrition is also more obvious [2].

According to the CDC [9], 42.4% of U.S. adults were obese in 2017-18, although the proportion was only 30.5% in 1999-2000 [10]. Notably, obesity is the most prevalent among people with lower educational levels and lower to middle income [3], who tend not to have the knowledge to design healthy diets or the money to hire nutritionists to do so [12].

Open Problem: the lack of nutrition knowledge and the challenge of knowing the complete knowledge base. Even though there has been an increased awareness of the nutrition health over the past years, the nutrition remains as a very specialized field, so that most people still cannot fully master the knowledge base or apply it correctly in the daily life. People might have read a number of articles, news and books on the healthy way for nutrition, but it is a challenge for everyone to consistently and correctly follow the best practice in any circumstances. Moreover, even for those people who has area of expertise in nutrition, it takes a good amount of time to evaluate healthy ingredients and meal every single day, as well as calculating calories. Calorie intake can vary by circumstances and add-on seasoning, thus it can be a fickle issue to tackle with. Therefore, this is why Health Diet is an essential tool to use.

Solution: a mobile system to automate analyzing the food type and calorie using AI and deep learning. As a result, we developed a program to identify foods and nutrition types from photos and calculate the approximate calorie intake. Firstly, this would allow users to identify what kinds of food/nutrition they lack and adjust their diets accordingly. Secondly, estimating the user's calorie intake helps the user control the amount of food he/she eats and thus better stay on her diet to keep weights off. The program's primary features include identifying food types from user input images and estimating the total calories. Secondary features include graphing daily intake curves for intuitive understandings into how well the user is following his/her diet and a "share" button for users to share their meals.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the educational impact in Section 5; Related work is discussed in Section 6. Finally, Section 7 gives the conclusion remarks, as well as pointing out the future work of this project.

## 2. CHALLENGES

The following challenges are identified and address in the context of this research project.

### 2.1. Challenge 1: Getting the Complete and Optimal Dataset

Even though there are a large number of food images available online, the key challenge is the unavailability of the labelled images. All the experiments will be conducted using the needed label data for training come from in order to correctly classify and verify various foods. Generally speaking, in the downloaded data image training set, the number of different categories is unbalanced. In addition, the image does not always come in square shape, so a number of dataset pre-processing will be required.

### 2.2. Challenge 2: Conducting the Experiment using the Limited Computational Power

The fundamental quality of the deep learning application always lies in the quality and quantities of the dataset. As we are growing the number of available images in our dataset, it also increases the challenges of conducting the experiments with the limited computational power. How to balance the training quality and the training time becomes a new challenge in the field of deep learning. The best solution not only achieves the highest accuracy, but also can get the accuracy within a shorter period of training time.

### 2.3. Challenge 3: Selecting the Correct and Optimal Deep Learning Models

When it comes to model selection and training tuning, challenges may arise as well. For example, if the model is too big, it will not be good for moving context. Besides, if the model itself may not be accurate. We have been seeing a growing number of deep learning models developed and turned from the academia in the recent years. Although the new models and examples provide a lot more options to tackle the proposed problem, it also generates new challenges to make the most optimal selection from the available algorithms. Generating the comprehensive accuracy result will take more efforts and the new models also required a bigger number of training images in the dataset.

## 3. SOLUTION

### 3.1. Overview of the Solution “Healthy Diet”

The app “Healthy Diet” is able to identify the food type from images, calculate food calories, allow the users to make healthy diet plans with existing ingredients at home, and realize the diet with healthy, attainable methods.

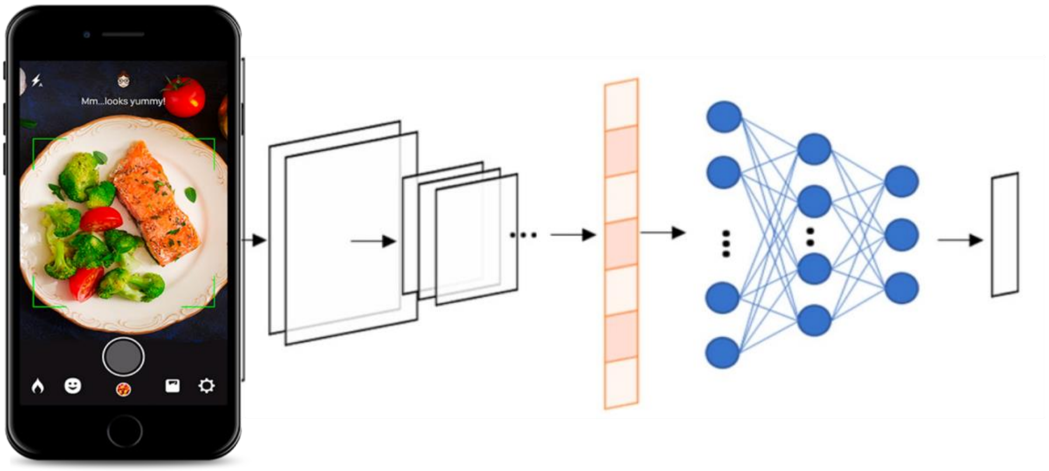


Figure 1. The Deep Learning-based Solution to Automate the Food Analysis

Figure 1 shows an overview of the solution. Health Diet can estimate the calories in the food. By having users taking pictures, Health Diet would identify the type of food through image recognition, retrieve the calorie data of the food from the database, and then use the approximate quantity of each food input by the user in order to multiply and estimate the calories in the food. The core part of the project is to classify images with neutral networks.

### 3.2. Features

In order to provide a decent user experience. A number of features have been proposed and implemented as follows:

a. *Meal Recommendations*

By identifying all the ingredients, Healthy Diet can provide users with healthy and delicious meal suggestions which match their taste. The app could guarantee a variety of healthy and low calories meals.

b. *Calculating Calories*

Users can record their own diet, generate a growth chart to count their own calorie intake over a period of time, which allows Health Diet record, analyze, and promote health eating habits.

c. *Share Eating Habits*

Users can share their own low-fat meals and track their own weight loss progress, which allow them to encourage each other to keep moving forward.

d. *Share Self-made Meal Plans*

Users can upload their self-made dishes, and after passing the review, the meal plans will be added to the recommended menu on the app and shared with other users.

## 4. EXPERIMENT

The major contribution of the work lies in the deep learning model training and the selection of the most optimal model and parameters to tackle this problem. The major research questions we are aiming to address are: 1) which deep learning model is the most accurate in classifying images for calorie analysis purpose; 2) how the number of epochs (number of times the program loops through the dataset; analogous the number of times a student reviews course materials) will impact the size of the model; 3) whether the decreased batch_size will increase model accuracy; 4) how is the model accuracy being affected by the changes of the number of epochs.

### 4.1. Deep Learning Models

In this experiment, multiple deep learning-based image classification models have been applied and compared, including the multiple layers of neural networks, convolutional neural networks, VGG, ResNet, EfficientNet, MobileNet. Among them, EfficientNet did not train successfully, because the log cannot be seen to determine the cause; at the same time, MobileNet has made progress, and the model recognition rate and size are acceptable, so we will focus on the optimization of MobileNet later.

### 4.2. Experiment Computing Environment

The training machine is mainly carried out on Tencent's smart titanium server, and it also uses its own computer and Kaggle for code debugging and simple model training.

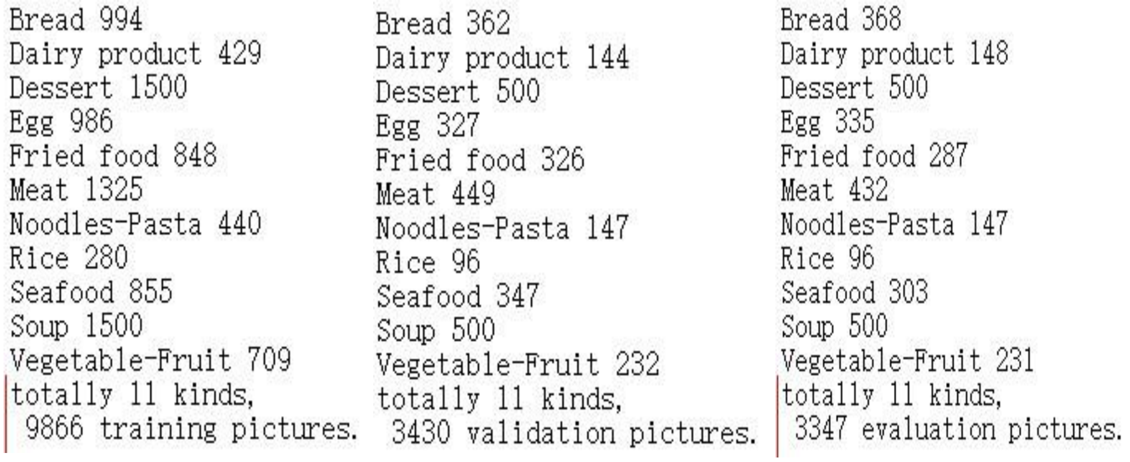
The initial multi-layer neural network and convolutional neural network were carried out on a personal computer without a GPU. Training for 50 or 60 epochs would take a few hours to run. Starting from VGG, after adjusting the script, it will run on the Tencent Smart Titanium server.

## 4.3. Training Dataset Analysis

The training data set are the 11-image-dataset that was downloaded from Kaggle. It classifies food into 11 categories:

"bread", " dairy products", "dessert ", "egg", "fried food", "meat ", "noodles/pasta noodles", "rice”, "seafood", "soup", "vegetable/fruit"

The compressed package size of the entire data set is 1.1G, divided into training/validation/evaluation three first-level sub-directories (number of files 9866/3430/3347), each sub-directory has a varying number of 11 food types second-level directories, in jpg format. The number of images for each food category is different. Taking training as an example, there are 1500 pictures in the Dessert/Soup secondary directory with the most, and 280 pictures in the Rice secondary directory with the least. The validation/evaluation are similar, the number of pictures of each food is less, and the ratio is roughly the same.



```
Bread 994
Dairy product 429
Dessert 1500
Egg 986
Fried food 848
Meat 1325
Noodles-Pasta 440
Rice 280
Seafood 855
Soup 1500
Vegetable-Fruit 709
totally 11 kinds,
 9866 training pictures.
```

```
Bread 362
Dairy product 144
Dessert 500
Egg 327
Fried food 326
Meat 449
Noodles-Pasta 147
Rice 96
Seafood 347
Soup 500
Vegetable-Fruit 232
totally 11 kinds,
 3430 validation pictures.
```

```
Bread 368
Dairy product 148
Dessert 500
Egg 335
Fried food 287
Meat 432
Noodles-Pasta 147
Rice 96
Seafood 303
Soup 500
Vegetable-Fruit 231
totally 11 kinds,
 3347 evaluation pictures.
```

Figure 2. Training Data Set of food

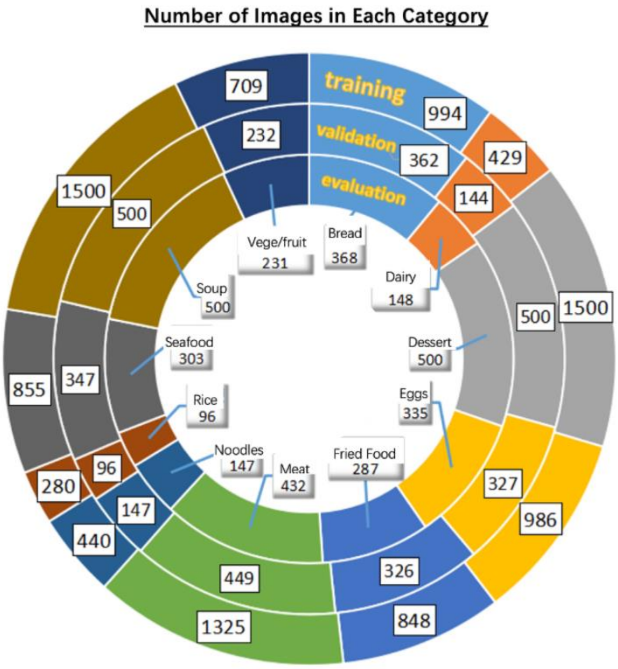


Figure 3. The Distribution of the Training Data Set Category

If the training data is extremely unbalanced, image preprocessing, under-sampling or oversampling or supplementation is often required, in order to prevent the model results from being biased towards the category of large data. There happened to be a period of test data comparison, but from the results, it is not very consistent. The test model uses 100 rounds of training results of ResNet18, and 15 pictures with known correct classifications are taken from each of the 11 types of food for prediction verification.

We have manually browsed and checked the catalogs. Most of the pictures (more than 90%) are square (the same number of pixels in length and width), a few (3-5%) are rectangles with different lengths and widths, and the few are the least. Exaggerated we have seen graphics with an aspect ratio close to 2:1 (for example, 512*288). The narrowest (length or width) pixels seen have 280+ pixels. In fact, this part of the data characteristics was not deliberately paid at the beginning. This was only noticed when thinking about how to improve the accuracy of the model in the later stage. Because improving the quality of training images is also an important means to improve accuracy.

## 4.4. Experiment Results

Figure 4 shows the overall performance of the selected machine learning models. The detailed results and comparison will be discussed in the following sub sections.

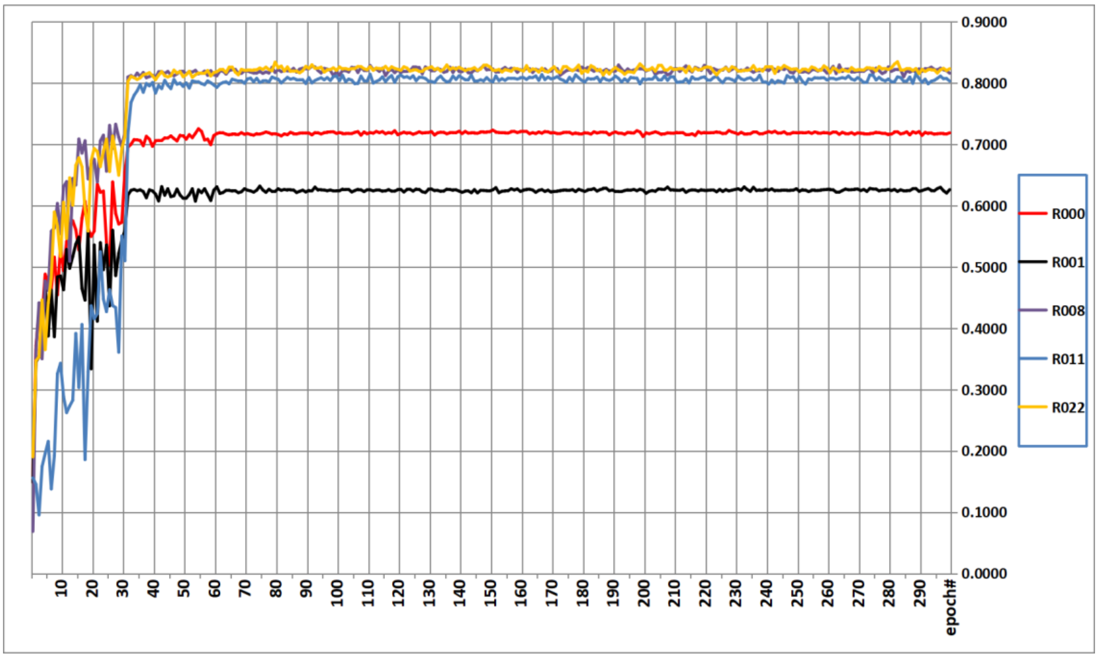


Figure 4. The Summary of the Machine Learning Model Performance

### 4.4.1. The Performance Comparison using MLP, CNN and VGG

Among these models, MLP and CNN [16][17] can still be trained on their own environment, and it takes about ten hours to train 60 epochs. It takes more than 8 hours to train 3 rounds of VGG, and it can only be trained on Tencent's smart titanium server, but the initial period may be improper use of the method, and no log can be seen. The log problem was solved in the later stage of ResNet and after MobileNet.

The problem encountered in the experiment is that after more than 30 epochs of CNN training, the training accuracy gradually increased to 93.2%+, while val_accuracy did not increase after 0.3x but slightly decreased. Checking the information relevant and thinking together, this should be an over-fitting phenomenon. After adding the Dropout() layer to the model, the phenomenon is alleviated, that is, val_accuracy no longer shows a downward trend, but after rising to a certain value (0.5), the small-range fluctuations no longer rise.

The corresponding VGG model can run normally on the smart titanium server, because the log feedback cannot be seen, and the training accuracy and val_accuracy data during the training process cannot be known. It is seen that its performance is better than CNN, at about 94.1%.

The test result using MLP reached an accuracy rate of 90.2%. For some types of food such ash desserts, only 7 predictions are correct, which is the bottom of the accuracy rate as fried food, less than 50%; with so few pictures of rice, it predicted 9 pictures correctly (60%), which is slightly lower than the overall accuracy. Based on our experiment, we believe the cause for this issue are: 1) the mainstream training configuration from the Internet is for reference and could contain some hidden parameter variance; 2) its conclusions often require certain preconditions (e.g., whether our data should be classified as “extremely unbalanced training data”), and there are also models that are not up-to-date; 3) the problem can also be resulted from the issues such as insufficient training and insufficient test sample size.

### 4.4.2. The Performance Analysis on ResNet Model

We tried ResNet18 [15], by following the given reference code and library documentations, and the initially tried image enhancement techniques, including horizontal flip, up and down translation, or left and right translation. Dynamic learning rate is used for the initial learning. The learning rate uses the initial value of 0.001 (or 0.01/0.0001, etc.), and shifts one decimal point to the right every 30 epochs. The size of the h5 model is reduced to 135M. The test recognition accuracy reached 68%, and it was consistent.

Based on the experiments, it can be found that by reducing the number of downsampling (that is, when certain links are convolved, the original model stripe=2 is changed to stripe=1), it improves the model recognition rate. Taking into account that it is necessary to reduce downsampling to avoid the image falling too early to 1*1, and our model is 224*224, which is the standard input of ImageNet, which may not be necessary. Therefore, according to the standard map structure, the standard ResNet18 model was restored.

We have tried to learn the ResNet50 structure by ourselves and try to rewrite it on the basis of the ResNet18 code provided, but it was not very successful. At the same time, we tried to modify the code of EfficientNet, but failed to generate the model. Thus, due to the successful debugging of MobileNet, I concentrated on the research and development of MobileNet.

### 4.4.3. The Performance Analysis on MobileNet Model

After the model was commissioned, the training was continued in units of 10 epochs. According to the log judgment, the optimal val_accuracy appeared at epoch=55, which was 72.52%, but at that time, the training was continued and saved once every 10 epochs, so the model was not saved. The model that has a chance to be saved in the future is 72.05% of epoch=218.

The size of the MobileNet [13][14] model is about 1/9 of the ResNet model. The h5 model is 15M, after the conversion is 6.8M, it is relatively faster to load on the mobile phone applet.

In order to obtain the historical peak value of val_accuracy (72.52% at epoch=55), we tried to start with the saved epoch=50 model and re-run 51-60, saving the model at every step. Because the image enhancement of the training parameters is inherently random, the parameters must be different during the actual operation. I have tried two scripts, one is that the log method is not changed at all (see the description of simplified log below), and the other is that the log writing is theoretically irrelevant to model training and we only care about val_accuracy, so the log is simplified. The result is that the highest version of the unmodified log has reached 71.61%, 71.55%, 71.41%, etc., and the simplified version of the log has 71.14% twice. However, the previous (referring to the 1-300#epoch run in multiple runs) has reached more than 71.8%, but the new 51-60#epoch in these two rounds has not reached. This test was only remembered today, and it was done temporarily.

In addition, we have also tried to add image enhancement parameters to val_data_gen, but the result is that val_accuracy has decreased (peak value is more than 60). It can be seen that the validation link and training use different pictures independently, so this change should not affect the improvement of the training effect itself, but affect the consistent standard of evaluation. Therefore, in other attempts, no image enhancement parameters were applied to val_data_gen.

The image enhancement in the train link adds the rotation_range=20 parameter, which is intended to randomly rotate the image by plus or minus 20 degrees. In effect, val_accuracy also drops, with a peak value of more than 60. We also tried to simplify the log. By setting the environment variable TF_CPP_MIN_LOG_LEVEL=3 and the fit_generator() parameter verbose=2, only the val_accuracy we care about is displayed. The effect is good, and the log is much more concise. Finally, the parameter workers=n in fit_generator() allows the maximum number of threads. Appropriate settings can increase the training speed. But in the process of gradually trying to increase this parameter, once workers=8 resulted in no log.

After reviewing what has been done in the some other related work [5][6][7], the MobileNet model can reach 90%+ after various skills blessings, but we only have 70%+. One of the reasons is the training images. For example, we did not conduct preprocessing work like checking and correcting images one by one. When we specify the length * width parameters when training the data, the machine will resize without considering the appropriate cropping or scaling. Another factor is that when resizing, it is best to reduce the original image, rather than to enlarge it. Visual inspection-can only rely on visual inspection first-the smallest length/width pixel value is also close to 300, so resize to 224*224.Papers in this format must not exceed twenty (20) pages in length. Papers should be submitted.

## 5. Educational Impact

When it comes to issues in K-12 educational settings, Health Diet can play an important role as well. According to Ruiz et al. [1], the children and adolescents with obesity has risen three times since 1970s, and those with server obesity has four times increment. It is also shown that children and adolescents with obesity are very likely to stay obese in their adulthood. It is an essential issue to treat obesity before adolescents enter adulthood, as obesity increase the likelihood to develop cardiovascular and mental illness.

Healthy Diet is widely applicable to this issue. If Healthy Diet is accepted by users in K-12 schooling, they will be able to create healthy meals on their own and learn how to keep their weight in a healthy range effortless. They do not have to worry about looking up random menus online or purchasing specific ingredients for a special diet. Health Diet is able to plan a health meal with the existing ingredients back at home, and it keeps tract of the calorie intake. Students can evaluate their progress just by accessing Health Diet.

On top of that, besides treating the obesity, Healthy Diet is also capable of ameliorating eating disorder. Social media has created unrealistic body images for adolescents, which has greatly jeopardized their mental health. It has already caused eating disorders to be a ubiquitous problem among teenagers. Adolescents are often eager to become skinnier, so they stop having any food several days in a row. Not only does this behavior harm physical health significantly, but it also causes body weight to fluctuate even more in the long term. If Health Diet is by adolescents, they will be able to lose their weight in a healthy matter, and maintain healthy lifestyle meanwhile. They can also evaluate their calories intake through viewing the sample menu to see if they are on the health level.

## 6. Related Work

Related works solve some problems such as inaccuracy, volume of food, and inability to recognize all food in an image [4]. Manika Puri, Zhiwei Zhu, Qian Yu, Ajay Divakaran, and Harpreet Sawhney [5] found a solution to old food intake assessment that suffer from inaccuracy or complex lab measurements. Their solution is to use a mobile phone to capture images of foods, recognize food types, estimate their respective volumes and finally return quantitative nutritional information. Yuji Matsuda, Hajime Hoashi, and Keiji Yanai [6] proposed a two-step method to recognize all the food in multiple-food images. Unlike these related works, we propose to apply these methods into developing a solution that helps people log their food and calories and ultimately helping them form a habit of recording their meals.

Liu et al. [7] applied an edge computing technique to apply deep learning on food recognition. The major difference between their work and our is that they focuses on identifying the food type only, while we are also analyzing the calorie amount. Similarly, Pouladzadeh et al. also applied the deep learning technique in the same domain, but their work targeted on trying to identifying multiple food images in the same image [8].

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a mobile application system to automate the analysis of food calories using deep learning and neutral network. According to the experiments we have conducted, the standard MobileNet can achieve a recognition accuracy of 76.69%, and after adding the two technologies of "reduction downsampling" and "data augmentation", it can achieve an astonishing 92.10% accuracy. For the food11 MobileNet model we trained here, the data augmentation has been included. The reduction of downsampling is modified on the basis of the reference code, and 3 reduction measures have been passively introduced. Thus, our MobileNet model should be against 92%, but it is actually only 72%. We think that the live broadcast class teacher uses the cifar10 data set, which recognizes handwritten numbers. Compared with the current food category, the difficulty should be different, so the decline in recognition accuracy should be reasonable.

In terms of the number of the users, the later model training, many explorations and parameter adjustments only improved the accuracy of a few percentage points or a few tenths of a percentage point. In this demonstration, increasing the accuracy of these percentages may have little effect. However, in many projects with a huge user base (possibly hundreds of millions) such as Tencent, this gap of a few thousandths or a few ten thousandths will have a huge impact. Therefore, whether some improvement methods are worth the effort depends on the environment in which they are placed, the user base and the accuracy requirements.

The experiment conducted has been very rewarding. This prototype has systematically sorted out the knowledge system of mathematics and artificial intelligence for us, and has opened the door for more foundation in AI. Although there are a lot of confusions in the learning process, we have gained more. It lays the foundation for our further study and understands a system. When we have the opportunity to learn relevant knowledge systematically in the future, we will know the direction of learning better.

Regarding the future work, we will be mainly focusing on two major aspects: 1) how to apply reinforcement learning in this problem and verify its accuracy; 2) performing more experiments with the increased number of training dataset and check the influence of the training dataset number; 3) we also want to build a mobile application that allows users to simply take a picture and get the result promptly.

Regarding In regards to future work in education, we will need to concentrate on three main goals: 1) Ensuring students who are on diet are controlling their weight in a healthy manner; 2) navigating students who are diagnosed with eating disorder to have each meal with enough nutrition; 3) Planning healthy daily meal for students who are not so familiar with meal preparation.

## REFERENCES

[1] Ruiz, L. D., Zuelch, M. L., Dimitratos, S. M., &amp; Scherr, R. E. (2019). Adolescent obesity: Diet QUALITY, Psychosocial health, and Cardiometabolic risk factors. Nutrients, 12(1), 43. https://doi.org/10.3390/nu12010043

[2] Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017–2018. NCHS Data Brief, no 360. Hyattsville, MD: National Center for Health Statistics. 2020 [https://www.cdc.gov/nchs/products/databriefs/db360.htm]

[3] Centers for Disease Control and Prevention. (2020, August 17). Losing weight. Retrieved January 30, 2021, from https://www.cdc.gov/healthyweight/losing_weight/index.html

[4] Kawano, Y., & Yanai, K. (2013). Real-time mobile food recognition system. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-7).

[5] Puri, M., Zhu, Z., Yu, Q., Divakaran, A., & Sawhney, H. (2009, December). Recognition and volume estimation of food intake using a mobile device. In 2009 Workshop on Applications of Computer Vision (WACV) (pp. 1-8). IEEE.

[6] Matsuda, Y., Hoashi, H., & Yanai, K. (2012, July). Recognition of multiple-food images by detecting candidate regions. In 2012 IEEE International Conference on Multimedia and Expo (pp. 25-30). IEEE.

[7] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Yunsheng, M., Chen, S. and Hou, P., 2017. A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. IEEE Transactions on Services Computing, 11(2), pp.249-261.

[8] Pouladzadeh, Parisa, and Shervin Shirmohammadi. "Mobile multi-food recognition using deep learning." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 13, no. 3s (2017): 1-21.

[9] Bray, George A., ed. "Obesity in America: A conference." (1979).

[10] Oliver, J. Eric, and Taeku Lee. "Public opinion and the politics of obesity in America." Journal of health politics, policy and law 30, no. 5 (2005): 923-954.

[11] Bean, Melanie K., Karen Stewart, and Mary Ellen Olbrisch. "Obesity in America: implications for clinical and health psychologists." Journal of Clinical Psychology in Medical Settings 15, no. 3 (2008): 214-224.

[12] Eagle, Taylor F., Anne Sheetz, Roopa Gurm, Alan C. Woodward, Eva Kline-Rogers, Robert Leibowitz, Jean DuRussel-Weston et al. "Understanding childhood obesity in America: linkages between household income, community resources, and children's behaviors." American heart journal 163, no. 5 (2012): 836-843.

[13] Qin, Zheng, Zhaoning Zhang, Xiaotao Chen, Changjian Wang, and Yuxing Peng. "Fd-mobilenet: Improved mobilenet with a fast downsampling strategy." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1363-1367. IEEE, 2018.

[14] Wang, Wei, Yutao Li, Ting Zou, Xin Wang, Jieyu You, and Yanhong Luo. "A novel image classification approach via dense-MobileNet models." Mobile Information Systems 2020 (2020).

[15] Bethge, Joseph, Christian Bartz, Haojin Yang, Ying Chen, and Christoph Meinel. "MeliusNet: Can binary neural networks achieve mobilenet-level accuracy?." arXiv preprint arXiv:2001.05936 (2020).

[16] Lei, Xinyu, Hongguang Pan, and Xiangdong Huang. "A dilated CNN model for image classification." IEEE Access 7 (2019): 124087-124095.

[17] Liu, Zhe, Wei Qi Yan, and Mee Loong Yang. "Image denoising based on a CNN model." In 2018 4th International Conference on Control, Automation and Robotics (ICCAR), pp. 389-393. IEEE, 2018.

© 2021 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

# APPLYING AI AND BIG DATA FOR SENSITIVE OPERATIONS AND DISASTER MANAGEMENT

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China.

## ABSTRACT

*Artificial intelligence has been a buzz word that is impacting every industry in the world. With the rise of such advanced technology, there will be always a question regarding its impact on our social life, environment and economy thus impacting all efforts exerted towards sustainable development. In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets for different industries and business operations. Numerous use cases have shown that AI can ensure an effective supply of information to citizens, users and customers in times of crisis. This paper aims to analyse some of the different methods and scenario which can be applied to AI and big data, as well as the opportunities provided by the application in various sensitive operations and disaster management.*

## KEYWORDS

*Artificial Intelligence, Big Data, Sensitive Operations, Disaster Management*

## 1. INTRODUCTION

Artificial intelligence (AI) is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think. AI is accomplished by studying how human brain thinks, and how people learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems [1]. AI is a science and innovation based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering. A major thrust of AI is in the development of computer functions associated with human intelligence, for example, reasoning, learning, and problem solving. Out of the following areas, one or multiple areas can contribute to build an intelligent system [2]. This paper aims to analyse some of the use of big data for the AI development and its applications in various sensitive business operations and disaster management.

## 2. WHAT IS BIG DATA

The Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. A buzzword that is used to describe immense volumes of data, unstructured, structured and semi-structured, big data can inundate a business on a day-to-day basis. Big data is used to analyse

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 349-356, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111429

insights, which can lead to better decisions and strategic business moves [3]. The definition of big data: “Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” The characteristics of Big Data are commonly referred to as the four Vs:

**Volume of Big Data**

The volume of data refers to the size of the data sets that need to be analysed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Asia.

**Velocity of Big Data**

Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Instagram messages or Wechat posts.

**Variety of Big Data**

Variety makes Big Data really big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.

**Veracity of Big Data**

Veracity refers to the quality of the data that is being analysed. High veracity data has many records that are valuable to analyse and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labelled Big Data [4].

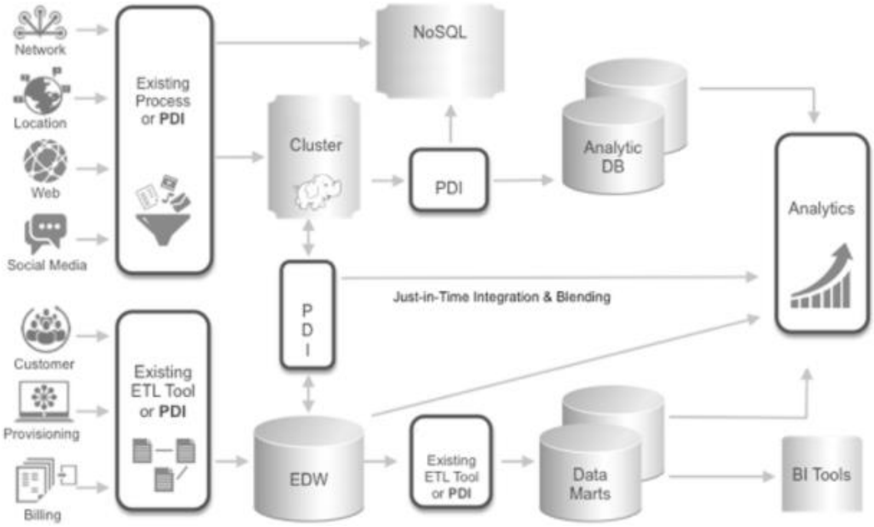


Figure 1. Big Data Architecture. (arccil.com)

## 2.1. Types of Big Data

There are 3 types of big data; unstructured data, structured data and semi-structured data.

**Unstructured data:**

Any data with unknown form or the structure is classified as unstructured data.

**Structured data:**

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

**Semi-structured data:**

Semi-structured data can contain both the forms of data.

Dealing with unstructured and structured data, data science is a field that comprises everything that is related to data cleansing, preparation, and analysis. Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data [5]. This umbrella term includes various techniques that are used when extracting insights and information from data.

**Big data benefits:**

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data, which means a completely different approach to tackling problems.

## 2.2. What is Big Data Analytics

Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models [6] [7]. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

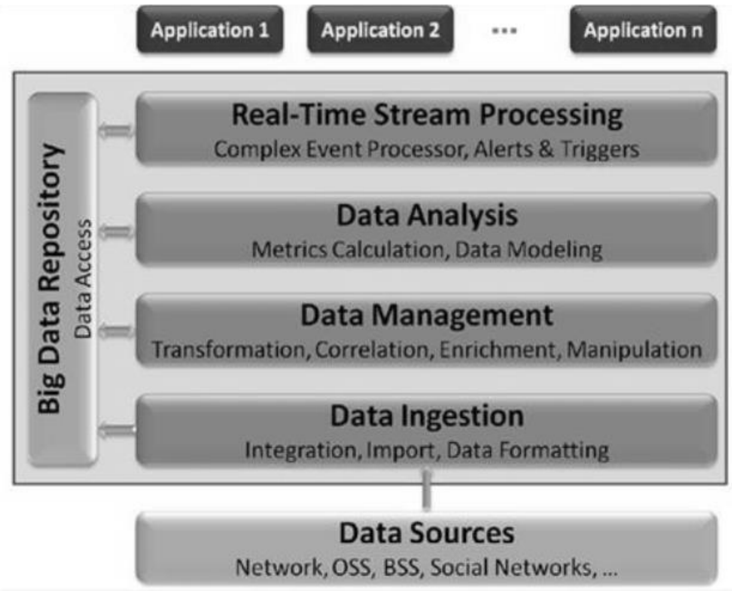

Application 1
Application 2
...
Application n
Big Data Repository
Data Access
Real-Time Stream Processing
Complex Event Processor, Alerts & Triggers
Data Analysis
Metrics Calculation, Data Modeling
Data Management
Transformation, Correlation, Enrichment, Manipulation
Data Ingestion
Integration, Import, Data Formatting
Data Sources
Network, OSS, BSS, Social Networks, ...

Figure 2. Big Data Analytics Architecture.

# 3. USING AI IN SENSITIVE BUSINESS OPERATIONS

The artificial intelligence rules define the way the online learning system assigned learning materials and exercises for the learner to follow [8]. These are the basic rules which we have carry out in our experiments, in which we find it effective in improving the learners understanding.

## 3.1. Financial Industry

Artificial intelligence (AI), along with other financial technology (fintech) innovations, are significantly changing the ways that financial business are being run, especially in the fields like trading and insurance, leading the traditional financial industry into a new era [9].

### Robots Replacing Humans

Back in 2000, Goldman Sach's New York headquarters employed 600 traders, buying and selling stock on the orders of the investment bank's clients. Today there are just two equity traders left, as automated trading programs have taken over the rest of the work. Meanwhile, BlackRock, the world's biggest money manager, also cut more than 40 jobs earlier this year, replacing some of its human portfolio managers with artificially intelligent, computerized stock- trading algorithms. Those two big companies are not the only financial institutions replacing human jobs with robots.

By 2025, AI technologies will reduce employees in the capital markets by 230,000 people worldwide, according to a report by the financial services consultancy Opimas [10].

Big new frontiers are only just beginning to opening up in fintech from AI, block chain and robotics to biometrics, augmented reality and cybersecurity. Among all the fintech innovations, the prospect of the block chain has the highest expectation. The block chain will change the way people store information, which is real, spreading fast and cross-border, and its 'de-centric' feature will allow everyone to know what other people are doing. The application of block chain in finance will once again bring about a revolutionary impact on the industry, just like AI does.

## 3.2. Health Industry

The Artificial intelligence (AI) is reshaping operations across industries. Arguably, healthcare is where these changes are poised to make the biggest impact – optimizing uptime and availability of the treatment solutions. Using AI-powered tools capable of processing large amounts of data and making real-time recommendations, healthcare organizations are learning they can reduce administrative waste in a number of areas, from medical equipment maintenance to hospital bed assignments [11].

Artificial intelligence is reinventing and reinvigorating modern healthcare through technologies that can predict, comprehend, learn and act. The ability of AI to transform clinical care has received widespread attention, but the technology’s potential extends beyond patient care to processes across the spectrum of healthcare operations. In healthcare and other industries that depend on reliable equipment performance, few things are more disruptive than unexpected outages. These unplanned stops create costly emergency situations, such as extended downtime, rush delivery of parts and overtime to repair the equipment.

Facing pressure to improve profitability and efficiency, many healthcare organizations are turning to emerging technologies like AI and big data analytics to improve upon existing maintenance operations. Until recently, maintenance typically involved either reacting to an unexpected problem or adhering to a preventive maintenance schedule, which can sometimes result in unnecessary maintenance. line.

## 3.3. Manufacturing Industry

AI is core to manufacturing's real-time future. Real-time monitoring provides many benefits, including troubleshooting production bottlenecks, tracking scrap rates, meeting customer delivery dates, and more. It's an excellent source of contextually relevant data that can be used for training machine learning models. Supervised and unsupervised machine learning algorithms can interpret multiple production shifts' real-time data in seconds and discover previously unknown processes, products, and workflow patterns [12].

The manufacturing industry has exploited the use of AI technology, and in particular knowledge-based systems, throughout the manufacturing lifecycle. This has been motivated by the competitive challenge of improving quality while at the same time decreasing costs and reducing design and production time. Just-in-time manufacturing and simultaneous engineering have further required companies to focus on exploiting technology to improve manufacture planning and coordination, and on providing more intelligent processing in all aspects of manufacturing. The objective is to improve quality, to reduce costs, and to speed up the design and manufacturing process.

# 4. USING AI IN DISASTER MANAGEMENT

## 4.1. Extreme Weather Forecast

According to the UN Office for the Coordination of Human Affairs, in 2016 over 100 million people were affected by natural disasters including earthquakes, hurricanes and floods. Technology has a vital role to play in providing the appropriate situational awareness that then shapes practical, life-saving decisions for effective crisis management. These decisions may involve the evacuation of the most dangerous areas after an earthquake, or explore tactical options about how and where to position critical resources like medicine, food, clean water and shelter. Through utilising the data tweeted and texted by citizens in a crisis zone, rescuers have access to the knowledge needed to devise a strategy for immediate rescue attempts and for longer term help [13].

Issues can arise, however, due to the volume of available data, and high-quality filtering systems are needed to avoid using inaccurate data that could misdirect humanitarian aid, potentially wasting time, resources, and human trust in the system. Humanitarian responders may, understandably, question the specificity of information, therefore, building their trust and encouraging uptake of AI technology is a socially meaningful endeavour; without this, a system is unlikely to be adopted in the field. Machine learning, understood as the refinement of how AI 'learns' to use algorithms and other data, offers a solution to detecting key information taken from social media messages. Hence, researchers are focusing efforts on improving how the millions of messages are sifted by algorithms to overcome inaccuracy, ensuring that only the most important data is identified and shared.

## 4.2. Man-Made Environmental Disaster

The case of BP oil spill in 2010 provides an important example for understanding how these principles are valued by public opinion in a crisis situation, and how the communication actions by a corporation in this type of circumstances might have long-term effect on the brand image of the organization. On April 20, 2010, a BP's Deepwater Horizon oil rig exploded, causing what has been called the worst environmental disaster in U.S. history and taking the lives of 11 rig workers. For 87 straight days, oil and methane gas spewed from an uncapped well-head, 1 mile below the surface of the ocean. The federal government estimated 4.2 million barrels of oil spilled into the Gulf of Mexico [14].

The accumulation of unsafe supervisory action had resulted in risk levels substantially increasing. Not only were risks increasing, but they were also incrementally becoming more aggressive in nature. For instance, one of the first acts of unsafe supervision is illustrated when BP neglected its responsibility of ensuring safety protocols were carried out after the completion of the Macondo Well. This was a major mistake on BP's part, violating safety protocols which may have identified the issues present with the cementing of the well. Should these issues have been identified sooner, the likeliness of the crisis happening would potentially be slim. In addition to this, there was also very little supervision during and after works were carried out. This can be attributed to the aforementioned organisational restructuring which created much confusion regarding who was accountable for the assurance of safety [15].

## 4.3. Natural Disaster

Researchers have found that AI can be used to predict natural disasters. With enormous amounts of good quality datasets, AI can predict the occurrence of numerous natural disasters, which can be the difference between life and death for thousands of people [16]. Some of the natural disasters that can be predicted by AI are:

**Earthquakes:**

AI systems can be trained with the help of seismic data to analyse the magnitude and patterns of earthquakes and predict the location of earthquakes and aftershocks.

**Floods:**

Various researchers and technology experts are developing AI-based applications with the help of rainfall records and flood simulations to predict and monitor flooding.

**Volcanic eruptions:**

AI-powered systems can accurately predict volcanic eruptions with the help of seismic data and geological information.

**Hurricanes:**

AI can use satellite to predict and monitor the path and intensity of hurricanes and tornadoes.

## 5. CONCLUSIONS

The study is assessing new frameworks for effective prevention measures and how AI can fit in and foster the early warning process. So further experiments and understanding the interrelation between AI and big data, what frameworks and systems that worked, and how AI can impact on different business operations whether by introducing new innovations that foster crisis management learning process and early prevention measures. The study from various reviews show promising results in using AI to learn specific industry big data and further evaluation and research is in progress.

## REFERENCES

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, (2015). Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8), pp.228- 232.

[2] A. Gandomi and M. Haider, (2015). Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2), pp.137-144.

[3] C. Lynch, (2008). Big data: How do your data grow?, Nature, 455, pp.28-29.

[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, (2015). Significance and challenges of big data research, Big Data Research, 2(2), pp.59-64.

[5] R. Kitchin, (2014). Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1).

[6] C. L. Philip, Q. Chen and C. Y. Zhang, (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275, pp.314-347.

[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, (2014). Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, (2014). On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285, pp.112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, (2014). Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1, pp.114-126.

[10] Xinhua China Daily, (18-Sep-2017). How is AI disrupting financial industry. http://www.chinadaily.com.cn/business/2017-09/18/content_32147126.htm

[11] Focus Elekta's Online Magazine, (2019). How AI is revolutionizing healthcare operations, https://focus.elekta.com/2019/10/how-ai-is-revolutionizing-healthcare-operations/

[12] Louis Columbus, (18-May-2020). 10 Ways AI is improving manufacturing in 2020, Forbes. https://www.forbes.com/sites/louiscolumbus/2020/05/18/10-ways-ai-is-improving- manufacturing-in-2020/?sh=3530e5d1e85a

[13] Kejriwal M. & Zhou P., (2019). SAVIZ: Interactive Exploration and Visualization of Situation Labeling Classifiers over Crisis Social Media Data, International Conference on Advances in Social Networks Analysis and Mining, Vancouver, Aug 27-30, pp705-708.

[14] National Commission, (2011). The Gulf Oil Disaster and the Future of Offshore Drilling.

[15] Dhaimaan Mahmud, (2019). Crisis Management Analysis of the BP Oil Spill, Birmingham Business School.

[16] Naveen Joshi, (15-Mar-2019). How AI can and will predict disasters, Forbes, https://www.forbes.com/sites/cognitiveworld/2019/03/15/how-ai-can-and-will-predict-disasters/?sh=7cddf40d5be2

## AUTHOR

**Prof. Yew Kee Wong** (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, Hewlett- Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

## AUTHOR INDEX