NLP Techniques and Applications

`

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 2nd International Conference on NLP Techniques and Applications (NLPTA 2021), November 27~28, 2021, Dubai, UAE
- 6th International Conference on Education (EDU 2021)
- 2nd International Conference on Data Science and Applications (DSA 2021)
- 2nd International Conference on Internet of Things & Embedded Systems (IoTE 2021)
- 12th International Conference on VLSI (VLSI 2021)
- 11th International Conference on Digital Image Processing and Pattern Recognition (DPPR 2021)
- 11th International Conference on Advances in Computing and Information Technology (ACITY 2021)
- 11th International Conference on Artificial Intelligence, Soft Computing and Applications (AIAA 2021)
- 8th International Conference on Computer Networks & Data Communications (CNDC 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: 978-1-925953-53-4
DOI: 10.5121/csit.2021.111901- 10.5121/csit.2021.111917

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 NLP Techniques and Applications (NLPTA 2021), November 27~28, 2021, Dubai, UAE, 6th International Conference on Education (EDU 2021), 2nd International Conference on Data Science and Applications (DSA 2021), 2nd International Conference on Internet of Things & Embedded Systems (IoTE 2021), 12th International Conference on VLSI (VLSI 2021), 11th International Conference on Digital Image Processing and Pattern Recognition (DPPR 2021), 11th International Conference on Advances in Computing and Information Technology (ACITY 2021), 11th International Conference on Artificial Intelligence, Soft Computing and Applications (AIAA 2021) and 8th International Conference on Computer Networks & Data Communications (CNDC 2021) was collocated with 2nd International Conference on NLP Techniques and Applications (NLPTA 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 NLPTA 2021, EDU 2021, DSA 2021, IoTE 2021, VLSI 2021, DPPR 2021, ACITY 2021, AIAA 2021 and CNDC 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, NLPTA 2021, EDU 2021, DSA 2021, IoTE 2021, VLSI 2021, DPPR 2021, ACITY 2021, AIAA 2021 and CNDC 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 NLPTA 2021, EDU 2021, DSA 2021, IoTE 2021, VLSI 2021, DPPR 2021, ACITY 2021, AIAA 2021 and CNDC 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

Abbasi, Islamic Azad University, Iran
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abdelhadi Assir, Hassan 1st University, Morocco
Abdelhak Merizig, University of Biskra, Algeria
Abdelkaher AIT Abdelouahad, Chouaib Doukkali University, Morocco
Abdellah Yousfi, University Mohammed V, Morocco
Abdellatif Bouzid-Daho, University of Tizi ouzou, Algeria
Abdullah, Adigrat University, Africa
Abhay Kumar Agarwal, Kamla Nehru Institute of Technology, India
Abhilash, Cyrielle Castle, India
Abir Messaoudi, University of Quebec, Canada
Addisson Salazar, Universitat Politècnica de València, Spain
Adnan Rawashdeh, Yarmouk University, Jordan
Ahmed Farouk AbdelGawad, Zagazig University, Egypt
Ahmed Refaat Ragab, Carlos III University, Spain
Ajit Singh, Patna University, India
Akhil Gupta, Lovely Professional University, India
Akhil Jabbar, Vardhaman College of Engineering, India
Alexander Gelbukh, Instituto Politécnico Nacional, Mexico
Ali A. Al-Zuky, Mustansiriyah University, Iraq
Ali A. Amer, Taiz University, Yemen
Ali Karkeh Abadi, University of Tehran, Iran
Alia Karim AbdulHassan, University of technology iraq, Iraq
Alireza Valipour Baboli, University Technical and Vocational, Babol, Iran
Alper Ugur, Pamukkale University, Turkey
Amando P. Singun, Higher College of Technology, Oman
Amel BorgI, Université de Tunis El Manar, Tunisia
Amel Ourici, Badji Mokhtar University of Annaba, Algeria
Amine Berqia, ENSIAS-Mohammed V University in Rabat, Morocco
Anamika Ahirwar, Jayoti Vidyapeeth Women's University, India
Anand nayyar, Duy Ten University, Viet Nam
Anandakumar Haldorai, Sri Eshwar College of Engineering, India
Anchit Bijalwan, Arba Minch University, Ethiopia
Anirban Banik, National Institute Of Technology Agartala, India
António Abreu, ISEL- Polytechnic Institute of Lisbon, Portugal
Arcely Perez-Napalit, Holy Angel University, Philippines
Arman Roohi, University of Central Florida, FL, USA
Asadollah Shahbahrami, University of Guilan, Iran
Ashraf Elnagar, College of Computing and Informatics, UAE
Atena Abdi, Amirkabir University of Technology, Iran
Attila Kertesz, University of Szeged, Hungary
Awais Khan, Sogang University, South Korea
Ayush Dogra, CSIR-CSIO, India
Azeddine WAHBI, University Hassan II, Morocco

`

B D C N Prasad,	V R Siddhartha Engineering college, India
Barbaros Preveze,	Cankaya University, Turkey
Behrouz Gordan,	Islamic Azad University, Iran
Bola Oladejo,	University of Ibadan, Nigeria
Bouchra Marzak,	Hassan II University, Morocco
Boukari Nassim,	skikda university, Algeria
Brahim Lejdel,	University of El-Oued, Algeria
Ch.Satyananda Reddy,	Andhra University, India
Cheng Siong Chin,	Newcastle University, Singapore
Ching-Nung Yang,	National Dong Hwa University, Taiwan
Claudio Schifanella,	University of Turin, Italy
Dadmehr Rahbari,	Tallinn University of Technology, Estonia
Daming Feng,	CGG research and development team, USA
Dan Wan,	Hunan Normal University, China
Daniel Imanah,	CEO- Zudan Electric Ltd, Nigeria
Dário Ferreira,	University of Beira Interior, Portugal
Debbat Fatima,	Mascara University, Algeria
Debotosh Bhattacharjee,	Jadavpur University, India
Deniz Kavi,	The Koç School, Turkey
Devlina Adhikari,	Indian Institute of Technology Kharagpur, India
Dimitris Kanellopoulos,	University of Patras, Greece
Dinesh Reddy,	CSE SRM, India
Djordje Cantrak,	University of Belgrade, Serbia
Domenico Rotondi,	FINCONS SpA, Italy
E.L Pradeesh,	Bannari Amman Institute of Technology, India
Eduardo O. Sanchez,	Universidad Politécnica de Madrid, Spain
EL Murabet Amina,	Abdelmalek Essaadi University, Morocco
Elnaz Pashaei,	Istanbul Aydin University, Turkey
Elżbieta Macioszek,	Silesian University of Technology, Poland
Endre Pap,	Singidunum University, Serbia
Essam Sourour,	Alexandria University, Egypt
EZ-Zahout Abderrahmane,	Mohamed V University, Morocco
F. Abbasi,	Islamic Azad University, Iran
Faisal Imran,	King Mongkut University, Thailand
Farhi Marir,	Zayed University, UAE
Farid Kheiri ,	Technical Vocational Training Organisation, Iran
Farouq Saber Al-Shibli,	Philadelphia University, Jordan
Felix J. Garcia Clemente,	University of Murcia, Spain
Francesco Zirilli,	Sapienza Universita Roma, Italy
Francis Lugayizi,	North West University, South Africa
Frank Ibikunle,	Landmark University, Nigeria
Gayatri Mehta,	University of North Texas, USA
Ghasem Mirjalily,	Yazd University, Iran
Gholam Aghashirin,	Oakland University, USA
Gniewko Niedbała,	Poznan University of Life Sciences, Poland
Grigorios N. Beligiannis,	University of Patras, Greece
Grzegorz Sierpiński,	Silesian University of Technology, Poland
Guilong Liu,	Beijing Language and Culture University, China
Habil. Gabor Kiss,	Obuda University, Hungary
Hamed Taherdoost,	Canada West University, Malaysia
Hamid Ali Abed AL-Asadi,	Iraq University College, Iraq
Hamlich, Ensam,	Uh2c, Morocco

`

Haouassi Hichem,	Abbes Laghrour University Khenchela, Algeria
Hassan Hadi Saleh,	University of Diyala, Iraq
Hayfaa Abdulzahra Atee,	Middle Technical University, Iraq
Hedayat Omidvar,	National Iranian Gas Company, Iran
Hiromi Ban,	Sanjo City University, Japan
Hossein Khademolhosseini,	Islamic Azad University, Iran
Hussein Rahouma,	Minia University, Egypt
Ihab Zaqout,	Al-Azhar University, Palestine
Ijeoma Noella Ezeji,	University of Zululand, South Africa
Ilango Velchamy,	CMR Institute of Technology, India
Ilham Huseyinov,	Istanbul Aydin University, Turkey
Indranil Hatai,	Indian Institute of Technology Kharagpur, India
Intisar Al-Mejibli,	University of Essex, United Kingdom
Isa Maleki,	Islamic Azad University, Iran
Islam Tharwat Abdel Halim,	Nile University, Egypt
Israa Shaker Tawfic,	Ministry of Science and Technology, Iraq
Jabbar,	Vardhaman College of Engineering, India
Jackelou S. Mapa,	Saint Joseph INstitute of Technology, Philippines
Janusz Kacprzyk,	Polish Academy of Sciences, Poland
Jawad K. Ali,	University of Technology, Iraq
Jaydip Sen,	Praxis Business School, India
Jesuk Ko,	Universidad Mayor de San Andres (UMSA), Bolivia
Jibendu Sekhar Roy,	KIIT University, India
Jijesh J J,	Sri Venkateshwara College of Engineering, India
Jonah Lissner,	Israel Institute of Technology, Israel
Joydeep Bhattacharyya,	Intel, USA
K Lal Kishore,	JNTUH, India
K.L.Sudha,	Dayananda Sagar College of Engineering, India
Kadji Albert,	University of Ngaoundéré, Cameroon
Kamaraju M,	Gudlavalleru Engineering College, India
Kamel Jemai,	University of Gabes, Tunisia
Karim Mansour,	Salah Boubenider University, Algeria
Kassem Danach,	Islamic University of Lebanon, Lebanon
Katarzyna Rostek,	Warsaw University of Technology, Poland
Kazım Yildiz,	Marmara University, Turkey
Kazuyuki Matsumoto,	Tokushima University, Japan
Keivan Navi,	Shahid Beheshti University, Iran
Khader Mohammad,	Birzeit University, Palestine
Khurram Hameed,	Edith Cowan University, Australia
Klenilmar L. Dias,	Federal Institute of Amapa, Brazil
Labraoui Nabila,	University of Tlemcen, Algeria
Ljubomir Lazic,	UNION University Belgrade, Serbia
Loc Nguyen,	Independent scholar, Vietnam
Luisa Maria Arvide Cambra,	University of Almeria, Spain
M V Ramana Murthy,	Osmania University, India
M.A. Jabbar,	Vardhaman College of Engineering, India
M.K.Marichelvam,	Mepco Schlenk Engineering College, India
Malleswara Talla,	Concordia University, Canada
Mamoun Alazab,	Charles Darwin University, Australia
Manoj Kumar,	Vidya College of Engineering, India
Manuel Gericota,	Polytechnic of Porto, Portugal
Marcin Paprzycki,	Polish Academy of Sciences, Poland

`

Marichelvam,	Mepco Schlenk Engineering College, India
Mario Versaci,	DICEAM - Univ. Mediterranea, Italy
Mary Cherian,	Visvesvaraya Technical University, India
Masoud Asghari,	Urmia University, Iran
Maumita Bhattacharya,	Charles Sturt University, Australia
Mehdi Gheisari,	Islamic Azad University, Iran
Michail Kalogiannakis,	University of Crete, Greece
Min-Shiang Hwang,	Asia University, Taiwan
Mohamed Elhoseny,	Mansoura University, USA
Mohamed Fakir,	University sultan Moulay Slimane, Morocco
Mohamed Hamlich,	ENSAM, Morocco
Mohammad Qatawneh,	The University of Jordan, Jordan
Mohammad-Shahram Moin,	ICT Research Institute, Iran
Mohammed A.M. Salem,	German University, Egypt
Mohammed Aref Abdul Rasheed,	Dhofar University, Oman
Mohammed Atif,	Arab Open University, Kuwait
Mohammed Bouhorma,	Abdelmalek Essaadi University, Morocco
Mohammed Yacoab,	The New College Chennai, India
Mohankumar N,	Amrita Vishwa Vidyapeetham, India
Monika,	Chulalongkorn University, Thailand
Mozmin Ahmed,	East Indian School, India
Mueen Uddin,	Universiti Brunei Darussalam, Brunei
Muhammad Elrabaa,	KFUPM, Saudi Arabia
Mu-Song Chen,	Da-Yeh University, Taiwan
Nadia Abd-Alsabour,	Cairo University, Egypt
Nadine Akkari,	Lebanese University, Lebanon
Nancy Arya,	SGT University, India
Navaid Z. Rizvi,	Gautam Buddha University, India
Nazmus Saquib,	University of Manitoba, Nazmus
Nihar Athreyas,	University of Massachusetts, USA
Nitza Davidovitch,	Ariel University, Israel
Omid Mahdi Ebadati,	Kharazmi University, Iran
P Joseph Charles,	St. Joseph's College, India
Patrick Fiati,	Cape Coast Technical University, Ghana
Paulo Batista,	University of Évora, Portugal
Paulo Maciel,	Federal University of Pernambuco, Brazil
Pavel Loskot,	ZJU-UIUC Institute, China
Peiying Zhang,	China University of Petroleum, China
Periola Ayodele,	Bells University of Technology, Nigeria
Pr. Smain Femmam,	UHA University, France
R.Arthi,	SRM Institute of Science and Technology, India
Raimundas Savukynas,	Vilnius University, Lithuania
Rajeev Kanth,	University of Turku, Finland
Ramadan Elaiess,	University of Benghazi, Libya
Rami Raba,	Al azhar University, Palestine
Ramtin Mohammadi-Zand,	University of Central Florida, USA
Rao Li,	University of South Carolina Aiken, USA
Rituparna Datta,	University of South Alabama, USA
S.Ganapathy,	Vellore Institute of Technology, India
S.Vijayarani,	Bharathiar University, India
Saad Aljanabi,	Alhikma College University, Iraq
Saadia Driss,	University Hassan, Morocco

`

Sabah Suhail,	University of Tartu, Estonia
Sachin Kumar,	Kyungpook National University, South Korea
Sadeque Reza Khan,	National Institute of Technology, India
Sadique Shaikh,	KYDSC Trust's, India
Safawi Abdul Rahman,	Universiti Teknologi MARA, Malaysia
Sai Kumar T,	CMR Technical Campus, India
Said Agoujil,	University of Moulay Ismail Meknes, Morocco
Said Nouh,	Hassan II university of Casablanca, Morocco
Sebastian Floerecke,	University of Passau, Germany
Seyed Mahmood Hashemi,	KAR University, Iran
Shahid Ali,	AGI Education Ltd, New Zealand
Shahram Babaie,	Islamic Azad University, Iran
Shamneesh Sharma,	Poornima University, India
Shashikant Patil,	SVKMs NMIMS Mumbai, India
Shervan Fekri-Ershad,	Islamic Azad university of Iran, Iran
Shilpa Gite,	Symbiosis Institute of Technology, India
Shing-Tai Pan,	National University of Kaohsiung, Taiwan
Shin-Jer Yang,	Soochow University, Taiwan
Shiva Asadianfam,	Islamic Azad University, Iran
Shonkh Shuvro,	IIEST, India
Shrikant Tiwari,	Shrishankaracharya Technical Campus, India
Shubham Sharma,	University of Regina, Canada
Siarry Patrick,	Universite Paris-Est Creteil, France
Sidi Mohammed Meriah,	University of Tlemcen, Algeria
Sidnei José Buso,	INEP, Brazil
Sikandar Ali,	China University of Petroleum, Beijing, China
Siyue Wang,	Northeastern University, United States
Smain Femmam,	UHA University France, France
Souhila Silmi,	Wireless Network, Algeria
Sourav Sen,	Upstart Network Inc, USA
Sridhar Iyer,	S. G. Balekundri Institute of Technology, India
Subarna Shakya,	Tribhuvan University, Nepal
Subhendu kumar Pani,	Krupajal Computer Academy, India
Subil Abraham,	IBM, USA
Suhad Faisal Behadili,	University of Baghdad, Iraq
Sun-yuan,	National Cheng Kung University, Taiwan
Surabhi Srivastava,	University of KwaZulu-Natal, South Africa
swarnalatha p,	Vellore institute of technology, India
Thulani Phakathi,	North-West University, South Africa
Titas De,	Senior Data Scientist, India
U. Srinivasulu Reddy,	National Institute of Technology, India
Ulhas B Shinde,	CSMSS Chh. Shahu College of Engineering, India
Vahideh Hayyolalam,	Koç University, Turkey
Valerianus Hashiyana,	University of Namibia, Namibia
Vinod Pangracious,	Sorbonne University Paris, France
Wahbi Azeddine,	Hassan II University, Morocco
Wei Cai,	Qualcomm, USA
Wenyuan Zhang,	Southeast University, China
William R Simpson,	Institute for Defense Analyses, USA
Xiao Wang,	Amazon, USA
Xiao-Zhi Gao,	University of Eastern Finland, Finland
Xinrong Hu,	Wuhan Textile University, China

`

Xu Wang,	University of Technology Sydney, Australia
Xue Wu,	Tsinghua University, China
Yakoop Razzaz Hamoud Qasim,	Taiz University, Yemen
Yanyang Lu,	Donghua University, China
Youye Xie,	Colorado School of Mines, USA
Yuan Tian,	Nanjing Institute of Technology, China
Yu-Chen Hu,	Providence University, Taiwan
Yuping Fan,	Illinois Institute of Technology, USA
Yuping Yan,	ELTE, Hungary
Zakaria Kurdi,	University of Lynchburg, USA
Zhu Wang,	Sany Heavy Industry Corp, China
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)

`

2nd International Conference on NLP Techniques and Applications (NLPTA 2021)

6th International Conference on Education (EDU 2021)

2nd International Conference on Data Science and Applications (DSA 2021)

2nd International Conference on Internet of Things & Embedded Systems (IoTE 2021)

12th International Conference on VLSI (VLSI 2021)

11th International Conference on Digital Image Processing and Pattern Recognition (DPPR 2021)

`

11th International Conference on Advances in Computing and Information Technology (ACITY 2021)

11th International Conference on Artificial Intelligence, Soft Computing and Applications (AIAA 2021)

8th International Conference on Computer Networks & Data Communications (CNDC 2021)

AUTOMATED CHINESE ESSAY SCORING USING PRE-TRAINED LANGUAGE MODELS

Lulu Dong[1, 2], Lin Li[1, 2], HongChao Ma[3] and YeLing Liang[1, 2]

[1]The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai, Xi Ning, China
[2]Department of Computer Science, Qinghai Normal University, Xi Ning, China
[3]Beijing Language and Culture University, Beijing, China

ABSTRACT

Automated Essay Scoring (AES) aims to assign a proper score to an essay written by a given prompt, which is a significant application of Natural Language Processing (NLP) in the education area. In this work, we focus on solving the Chinese AES problem by Pre-trained Language Models (PLMs) including state-of-the-art PLMs BERT and ERNIE. A Chinese essay dataset has been built up in this work, by which we conduct extensive AES experiments. Our PLMs-based AES models acquire 68.70% in Quadratic Weighted Kappa (QWK), which out-perform classic feature-based linear regression AES model. The results show that our methods effectively alleviate the dependence on manual features and improve the portability of AES models. Furthermore, we acquire well-performed AES models with a limited scale of the dataset, which solves the lack of datasets in Chinese AES.

KEYWORDS

Chinese Automated Essay Scoring, Neural Network, Pre-trained Language Model, Quadratic Weighted Kappa.

1. INTRODUCTION

Writing is a measure of language learners meta-cognitive and linguistic abilities, thus Chinese writing draws increasing attention from learners. With the boom in learning Chinese all over the world, Chinese essay scoring becomes a challenge for both Chinese teaching and testing. It is not only because scoring essays is a time and labor-consuming task, but also different human raters have divergence on the same essays. AES is an effective and efficient substitution for human raters by assigning a holistic score to an essay. AES is a reasonable approach to alleviate the conflict between the increasing number of Chinese essays and the lack of human raters is helpful to reduce subjectivity in human scoring [1]. Classic feature-based AES systems have succeed for English AES like PEG [2], IEA [3], E-rater [4, 5], and BETSY [6]. The performance of a classic AES system is largely determined by its feature set, however, building a high-quality set is a time-consuming and laborious task. Furthermore, it is also a challenge even for experts to take all key scoring aspects into consideration. To reduce the dependency of AES systems on manually building feature sets, Neural Networks such as convolution and recursive neural networks have been introduced into the AES task [7]. The Neural Network-based AES approach avoids complex feature engineering but it is a corpus-greedy method. That is, a large scale of essay corpus is the prerequisite for acquiring a well-performed AES system.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 01-12, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111901

Language models succeed in many Natural Language processes communities because of their strong capability in language representation. The state-of-the-art PLMs [8] own powerful architecture and are trained on huge scale corpus by different pre-trained tasks. PLMs are successfully applied to complete AES tasks by researchers such as XLNET and BERT. Compared with feature-based and Neural Network-based AES methods, PLMs-based AES systems show a better agreement with human raters trained by the same scale of training corpora [9]. Currently, challenges for Chinese AES can be attributed to the lack of a powerful model and available large-scale corpus. To alleviate the difficulties, we propose to apply fine-tuned PLMs into the Chinese AES task in this work.

The rest of this paper is organized as follows. Section 2 provides an overview of related work in the literature. Section 3 gives a clear definition for the AES task and a detailed explanation for the evaluation metric used in this work. We provide the details of our approach in Section 4, present and analyze results in Section 5. Finally, we draw our conclusion in Section 6.

2. RELATED WORK

***Studies in English AES*.** The performance of sequence models like Long Short-Term Memory (LSTM) exceeds previous feature-based methods on AES tasks by its powerful capability of feature engineering and complex pattern encoding. Alikaniotis et al [10] applied Bidirectional LSTM (BiLSTM), based on which Fei and Yue [11] added convolution layer and pre-trained word embedding into their work. Attention mechanism has been used to AES models to alleviate gradient vanishing and long-distance dependency problem of sequence models. Fei et al [12] adopted attention into the AES task, by which the model can capture significant context and word-level information. With the success of the pre-trained language model in many NLP tasks, Rodriguez et al [13] proposed an AES model based on BERT [14] and XLNET [15]. In general, Neural Networks achieve better agreement with human raters in the English AES task.

Studies in Chinese AES. A few AES systems have been proposed for different Chinese tests like HSK (Hanyu Shuiping Kaoshi) [16], MHK (Chinese Proficiency Test for Minorities in China) [17], and high school essays test [18]. These systems assign a holistic score with linear regression models based on various linguistic features. Kakkonen et al [18,19] proposed a linear regression model whose features are represented by Latent Semantic Analysis (LSA). Previous AES work in Chinese shows a medium correlation between predicted scores and manual scores, which can not fully satisfy the practical application. Powerful modeling approaches have been used to improve the performance of Chinese AES like Supported Vector Machine (SVM) and Back Propagation (BP) neural network. For instance, Ma et al [20] made a comparison study in Chinese AES between SVM and BP neural network. And their results show that BP neural network achieves a higher correlation with the human raters than SVM. Fu [21] et al proposed a hybrid AES model that combined Recurrent Neural Network (RNN) with BiLSTM.

Pre-trained Language Models. PLMs have made great achievements in a variety of NLP tasks, for instance, BERT(Bidirectional Encoder Representation from Transformers) performs better than native human speakers in some GLUE benchmark tasks. PLMs attract increasing interest from organizations and researchers, thus many PLMs are proposed like BERT [14], ERNIE [22], and GTP [8]. BERT is one of the most successful PLMs, which is built on Transformer architecture and adopts a bidirectional mechanism. RoBERTa [23], a variant of BERT, is trained by more training data and training batches than BERT. Ernie is trained by different pre-trained tasks and training strategies that are helpful to add word and phrase-level information into the model. In this work, we focus on three PLMs based on Transformer architecture, thus we use BERT, ERNIE, and RoBERTa to explore the automatic score of Chinese essays.

3. AUTOMATED ESSAY SCORING FOR CHINESE

In this section, we provide a clear definition of the AES task, introduce the dataset we employed, and elucidate the evaluation metric used for assessing the performance of our AES models proposed in this work.

3.1. Task Description

An essay can be evaluated from different aspects including relevance to prompt, the correctness of grammar, and usage of lexical. In this work, we focus on holistic essay scoring, that is, assign an essay an overall score according to its general quality. Thus, the AES model proposed in this work takes a raw essay as input and provides a holistic score as output. We formalize Chinese AES as a classification task, which adopts a 12-class classification. Each class represents a range of 5 points assigned by human raters from 40 to 95.

Our classification models aim to deduce accurate score labels by learning implicit and explicit features of essays. What shall be noted is that scores assigned by human raters are treated as the golden standard, thus our models concentrate on maximizing the agreement between predicted scores and human raters.

3.2. Chinese Essay Corpora

Large-scale Chinese essay corpora with corresponding scores are essential for Chinese AES study. The corpora used in this work contains 7,141 Chinese essays, all of which were written by Chinese as a second language learner. Statistics of our corpora are shown in Table 1, which shows that the size of our dataset is only a third of ASAP dataset. We used 60%, 20%, and 20% essays of each prompt as training, validation, and test dataset, respectively. The corresponding scores are treated as labels adopted in this work, which is assigned by experts in a fair and strict scoring process. Specifically, each essay is graded by two independent human raters and a third senior rater will be introduced when the first two raters cannot reach an agreement in terms of the score.

Table 1. Statistic of our Chinese Essays Dataset

Prompt	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sample	861	703	825	198	325	739	1330	518	687	955
Rating range	40-95	40-95	40-95	65-95	40-95	40-95	40-95	40-95	50-95	40-95

3.3. Evaluation metrics

AES systems used to be evaluated by a variety of statistical measures like Pearson correlation coefficient and Kappa [24]. Since the ASAP competition adopts Quadratic Weighted Kappa (QWK) as an evaluation metric, QWK becomes a widely used approach for AES task by several work [24]. We also take QWK as our evaluation method not only because it is a popular method to that ensure our result is comparable with other AES work, but also because its principle is more reasonable for evaluating AES models than other metrics.

QWK is capable of capturing disagreement between two raters by fusion matrix. The range of QWK is from 0 to 1, and K = 1 if the two raters achieve complete agreement with each other. The brief procedure of computing QWK of two groups of essay scores is defined as follows. Firstly,

the weight matrix W is built up as defined by formula 1, where i and j are the scores given by a human rater and an AES system respectively, N is the number of essay grades.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \qquad (1)$$

Then, we construct the matrix Q and the prediction matrix E, where $Q_{i,j}$ refers to the number of times an essay is graded as i by human raters and as j by the automatic score method. E is the outer product of the manual score vector and the AES score vector. Finally, the QWK score is computed as the formula 2:

$$K = 1 - \frac{\sum_{i,j} W_{i,j} Q_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \qquad (2)$$

4. AES MODELS FOR CHINESE

Due to the size limitations of our dataset, only can we use a limited scale of samples to training a model. Thus we employ the strong ability of PLMs in encoding linguistic information and learning about the complex relationship between essays and corresponding scores and achieve a better performance in the AES task. Firstly we build up a regression-based AES model as our baseline model, which is an explainable and stable benchmark for this work. And several PLMs have been proposed with the success of PLMs in many NLP areas like Machine Translation, etc. Then we explore how to apply three powerful Chinese PLMs including BERT, ERNIE, and RoBERTa into our task.

4.1. Baseline Model

The regression-based AES method shows relatively stable and simple performance in both English and Chinese [25]. In the ASAP competition, A regression-based AES model won third place in the ASAP competition [26], following which we build up our benchmark for this work with multi-level latent linguistic features. We comprehensively consider features that contribute to essay scoring, by which we acquired an integrated feature set for Chinese AES. In this section, we will introduce the feature set we built up for Chinese AES, and describe the construction and implementation of the linear regression model.

Features Set A regression-based AES model needs a high-quality feature set. However, feature selection depends on manual selection task is full of challenges due to it is hard for people to comprehensively consider all key information for the AES task. So we construct a latent semantic feature set consisting of three aspects of linguistic features including characters, words, and sentences as shown in appendix Table A1.

Character-level features According to the difficulty, Chinese characters are divided into four grades by HSK [27]. Intuitively, the level of character usage is an effective measure to evaluate writing skills. That is, an essay consisting of more various and high-level characters will be rated a higher score. Based on this assumption, we proposed 11 character features.

Word-level features The words choosing and applying largely determine the level of an essay, so we extend the feature set with eight different features about word usage like the number of tokens,

misused words, etc. Through these features, the baseline model is capable of evaluating a writer's level of Chinese basic vocabulary, complex words, and phrases.

Sentence-level features We also consider sentence-level features except for character and word features, that including the ratio of the number of clauses to the total number of sentences, the average sentence length, and the total count of sentence errors. The sentence features ensure the baseline model can make a reasonable prediction of a writer's ability in complex sentence patterns and grammar knowledge.

Regression-based Model A linear regression model is a statistical approach which can make reasonable prediction by learning linear relationship between independent variables and dependent variables [28]. Multiple linear regression was constructed as the baseline model according to the function $Y = aX + b$, where X refers to the multidimensional features as input, and Y is the score predicted by the linear regression model. The linear regression model Y = aX+b can be realized by the full connection layer. Through a full connection layer, the input X and output Y are connected and the parameters a and offset term b are allocated.

The mean square error (MSE) is taken as the loss function, the Root Mean Square Prop (RMSProp) is the optimization algorithm, and the extracted features are used as the linear regression model of dataset training. In the construction of the linear regression model, the normalization technology is used to normalize the extracted feature data. The data in different ranges are in the same distribution range, which can make the optimization algorithm have a faster convergence speed.

4.2. Our PLMs-based AES Models

Neural Networks require large-scale training data for learning complex patterns between essays and corresponding scores. To avoid the over-fitting problem [7], we explore how to effectively apply PLMs into the Chinese AES.

BERT is the abbreviation for Bidirectional Encoder Representations from Transformers [14], which is one of the most successful PLMs and outperforms human participants in many tasks of GLUE. BERT's success can be attributed to its effective language modeling approach and the bidirectional multi-layer transformer architecture. As a deep bidirectional transformer model, which generates a feature vector for each element (like a word) of the input sequence with consideration of its preceding and succeeding context [29]. In this paper, we use the BERT-base Chinese to our task, which includes an embedding layer, 12 encoder layers, and a pooling layer. The parameters are up to 110M.

BERT consists of the Encoder of Transformer model and the Transformer learns the representation of linguistic units in context by self-attention and full connection layer. When encoding a linguistic unit of an input sequence, the self-attention mechanism of BERT determines assigning how much attention to each unit. These three vectors are the result of multiplying an embedding vector by a matrix W, which is randomly initialized. The self-attention is defined as for formula 3.

$$\text{Self-}attention = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \qquad (3)$$

Self-attention calculates three new vectors, which are called a query, key, and value (Q, K, and V) respectively, namely $Q = W^Q X^T$, $K = W^K X^T$, $V = W^V X^T$. Calculate the score of self-attention, which determines how much attention we pay to the rest of the input sentence when we encode a word in a certain position. The calculation method of this fractional value is to do point multiplication between query and key and divide the result of point multiplication by a constant $\sqrt{d_k}$. Then a softmax function is used to get the weight, and the result is the correlation of each word to the word in the current position. Multiply the value from value and softmax, and the result is the value of self-attachment in the current node.

BERT is trained by two different word level and sentence level tasks. One task is MLM (Masked Language Modeling) which learns language distribution by recovering several randomly masked words in a sentence. The other is NSP (next sense prediction) that main purpose is to predict the sequence of two sentences. Except for employing knowledge learning in the pre-training stage, BERT shall be properly fine-tuned for a specific downstream task. In this paper, we fine-tune the parameters of middle layers to adjust the embedding of the input sequence and parameters of the prediction layer to improve BERT's performance in our task (more details about fine-tuning are in Section 5).

RoBERTa Researchers of Facebook and the University of Washington carefully investigate the effects of hyper-parameters and scale of training corpus on the performance of BERT. And the result shows that the training of BERT is insufficient, thus they propose RoBERTa [23] improve BERT from the following aspects: (1) extending the scale of training corpus into 160G; (2) increasing the number of parameters into eight thousand; (3) extending training processing by using 500 thousand training epoch.

RoBERTa proposes a full-sentences mechanism, which refers to the length of the input sequence that has been extended from two sentences to a fixed-length context (i.e. paragraphs or articles). And the static MLM of BERT has been replaced with the dynamic MLM, that is, the inputs are masked just before used to the training. By these strategies, RoBERTa outperforms in many NLP tasks. In this paper we use RoBERTa-Base-Chinese to complete AES.

ERNIE employs the architecture of BERT, whose effectiveness has been proved by many studies. The primary idea of BERT is that a powerful representation of language can be effectively learned by simple pre-training tasks and a huge scale of the corpus. ERNIE proposes that the language representation can be further improved by more informative pre-training tasks. Thus, ERNIE adopts three different level mask units (words, phrases, and name entities) in MLM to acquire more semantic information [22].

Based on this assumption, ERNIE uses DLM (Dialogue Language Model) to model query response dialogue structure, takes dialogue pair as input, introduces dialogue embedding to identify the role of dialogue, and uses dialogue response loss to learn the implicit relationship of dialogue, to further improve its semantic representation ability. ERNIE can potentially learn knowledge dependency and longer semantic dependency by unifying the mask to make the model more generalized. Experiments show that ERNIE achieves good results in some Chinese NLP tasks, which is related to its use of forum data for dialogue modeling, which the training corpus for ERNIE is multi-source like encyclopedia article and dialogue, it ensures ERNIE learn various language information distributed in different genre of the corpus.

4.3. Fine-tuning Strategy

When applying PLMs to downstream tasks, many difficulties need to be solved like catastrophic forgetting, which makes the model quickly forget what it learned before. In the work, we use three basic versions of the pre-training language model and try to use a variety of fine-tuning strategies to obtain better experimental results. In the process of fine-tuning, we mainly focus on the influence of sequence length, learning rate, and batch-size on the experimental results. In addition, the final optimization strategy and operation configuration parameters are shown in Table 2.

Table 2. Parameter configuration

Parameter	Value	Parameter	Value	Parameter	Value
optimizer	AdamWeightDecayStrategy	lr-scheduler	Linear decay	learning-rate	2e-5
weight-decay	0.01	Max-seq-len	510	batch-size	32
warmup-proportion	0.1				

5. EXPERIMENTS AND RESULTS

5.1. Experiments

In this work, we run intensive experiments to investigate the way how to solve the Chinese AES problems by different PLMs including BERT, ERNIE, and RoBERTa. To compare and analyze the performance of different PLMs, we utilized the same dataset in all experiments. Google Colab platform was used to execute our baseline model, and the Baidu PaddlePaddle platform was employed to training and testing our PLMs models.

5.2. Results and Analysis

We tentatively make use of CNN and RNN in Chinese AES, and our results show worse performance than our baseline system. We think the bad performance of CNN and RNN is because they cannot learn about the complex relationship between essays and scores with a limited size of corpora. Our results suggest that it is not practical to training a Neural Network-based AES model from scratch.

Table 3. QWK Results (%)

Model	1	2	3	4	5	6	7	8	9	10	Avg
WSP-T-FT[30]	86.30	58.60	49.50	56.70	-	-	-	-	-	-	62.8
Baseline	57.83	52.40	48.25	62.27	55.32	61.88	56.39	59.19	60.92	59.58	57.40
BERT	77.15	74.83	68.43	73.08	62.24	75.21	56.09	64.51	50.97	72.10	67.46
ERNIE	76.00	79.96	69.62	65.10	62.24	76.36	58.92	61.72	48.68	76.54	67.51
RoBERTa	80.31	80.17	71.29	66.37	55.62	75.38	60.28	63.92	55.96	77.70	68.70

QWK results are shown in Table 3 and we also evaluate our models by Pearson correlation (as shown in Table 4) to compare with previous work. QWK is widely adopted for evaluating AES, while the Pearson coefficient could reflect ranking consistency. The WSP-T-Finetune model was Song [30] adapts multi-stage Pre-training strategy cooperate on the attentional recurrent convolutional neural network with the essay written by a Chinese student. The result shows that the pre-training-based approach is effective for AES. The average QWK of our baseline model is

57.40% and all the three PLMs-based AES models outperform our baseline system in terms of QWK. RoBERTa achieves the best QWK comparing with BERT and ERNIE, whose improvement of QWK reaches 11.30%. The different PLMs-based AES models show very similar performance in the AES task. Our results suggest that fine-tuned PLMs-based AES model is a practical way for Chinese AES with the limitation of the scale of the corpus. In addition, our PLMs-based models also obviously performed better than the work of [30] in the Chinese AES task.

Table 4. Experiment result Pearson correlation (%)

Model	1	2	3	4	5	6	7	8	9	10	Avg
WSP-T-FT[30]	87.70	62.90	53.40	60.60	-	-	-	-	-	-	66.20
Baseline	62.37	58.07	55.54	60.55	57.93	66.64	56.09	61.37	63.80	63.84	60.62
BERT	82.07	76.30	70.02	74.83	62.48	78.53	59.04	71.91	56.12	75.74	70.70
ERNIE	81.40	80.57	70.56	66.04	63.01	79.39	60.68	67.23	50.22	79.61	69.87
RoBERTa	83.40	81.05	72.71	66.37	56.55	77.14	61.26	70.91	57.64	79.84	70.69

A post-hoc analysis has been done for investigating the different performances of our models on different prompts. As shown in Figure 1, RoBERTa outperforms other models on most prompts (8 of 10) except for prompts 4 and 5. We think the best performance of RoBERTa in this work can be attributed to its long contextual training strategy, that is, RoBERTa learns a better language representation by using long context information. RoBERTa only falls behind other models on 2 prompts, we think this is caused by the very small sample size (Table 1). This result reveals the key effect of training set size on RoBERTa, that is, RoBERTa performs better than other PLMs models with the proper amount of training samples in the AES task. This assumption also can be proved by the obvious QWK improvement of RoBERTa on prompts 1, 3, 7, and 10 with a larger number of samples.

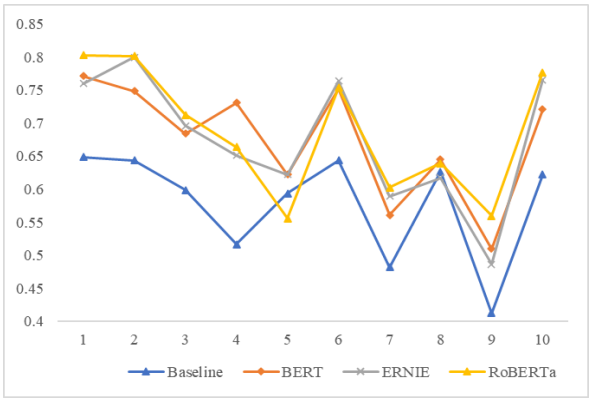


Figure 1. Results QWK under different prompts

6. CONCLUSION

In this paper, we build up a strong baseline system for the Chinese AES task by a linear regression model. Furthermore, we investigate how to apply PLMs into our AES task including BERT, RoBERTa, and ERNIE. By running intensive experiments on Chinese AES, we find that PLMs-based significantly outperform our baseline system. The results show that RoBERTa achieves 68.70% in QWK, which is 11.30% higher than the baseline. The designing and

performance of Chinese AES models are still limited to the size of our corpus, thus larger Chinese essay corpora will be helpful for further study in Chinese AES.

7. FUTURE WORK

In the future we will further analyze the internal structure of PLMs, a more reasonable fine-tuning strategy is adopted to further improve the effectiveness of the automatic scoring model. And we are interested in probe what features or traits are captured by PLMs for scoring. Furthermore, we will explore and make public the larger pre-training Chinese dataset with supervised labels or self-supervised learning strategies.

ACKNOWLEDGMENTS

The authors of this paper received support under Grant CTI2020B05 from Hankao International Education Technology (Beijing) Co., Ltd, which is gratefully acknowledged.

REFERENCES

[1] Balfour S P. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™[J]. Research & Practice in Assessment, 2013, 8: 40-48.
[2] Page E B. Grading essays by computer: Progress report[C]//Proceedings of the invitational Conference on Testing Problems. 1967.
[3] Foltz P W, Laham D, Landauer T K. The intelligent essay assessor: Applications to educational technology[J]. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1999, 1(2): 939-944.
[4] Attali Y, Burstein J. Automated essay scoring with e-rater® V. 2[J]. The Journal of Technology, Learning and Assessment, 2006, 4(3).
[5] Burstein J. The E-rater® scoring engine: Automated essay scoring with natural language processing[J]. 2003.
[6] Rudner L M, Liang T. Automated essay scoring using Bayes5 theorem[J]. The Journal of Technology, Learning and Assessment, 2002, 1(2).
[7] Taghipour K, Ng H T. A neural approach to automated essay scoring[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1882-1891.
[8] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020: 1-26.
[9] Mayfield E, Black A W. Should You Fine-Tune BERT for Automated Essay Scoring?[C]//Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2020: 151-162.
[10] Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks[J]. arXiv preprint arXiv:1606.04289, 2016.
[11] Dong F, Zhang Y. Automatic features for essay scoring-an empirical study[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1072-1077.
[12] Dong F, Zhang Y, Yang J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). 2017: 153-162.
[13] Rodriguez P U, Jafari A, Ormerod C M. Language models and automated essay scoring[J]. arXiv preprint arXiv:1909.09482, 2019.
[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[15] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. arXiv preprint arXiv:1906.08237, 2019.
[16] Liang Maocheng and Wen Quifang. Review and Enlightenment of foreign automatic scoring system for composition[J].Media in Foreign Language Instruction, 2007, 5: 18-24.
[17] Li Yanan. Automated Essay Scoring For Testing Chinese As A Second Language[D]. Beijing Language and Culture University.

[18] Cao Y, Yang C. Automated Chinese essay scoring with latent semantic analysis[J]. Examinations Research, 2007, 3(1): 63-71.
[19] Kakkonen T, Myller N, Timonen J, et al. Automatic essay grading with probabilistic latent semantic analysis[C]//Proceedings of the second workshop on Building Educational Applications Using NLP. 2005: 29-36.
[20] Ma Hongchao and Guo Li and Peng Hengli. Comparison of Automatic Scoring Effect of Writing Based on SVM and BP Neural Network[J]. Examination research, 2019, (5).
[21] Fu R, Wang D, Wang S, et al. Elegart sentence recognition for automated eassay scoring[J]. J. Chin. Inf. Process, 2018, 32(6): 88-97.
[22] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
[23] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[24] Yannakoudakis H, Cummins R. Evaluating the performance of automated text scoring systems[C]//Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. 2015: 213-223.
[25] Dikli S. An overview of automated scoring of essays[J]. The Journal of Technology, Learning and Assessment, 2006, 5(1).
[26] Thyagarajan A, Bhomick P K. Regression based Automated Essay Scoring[J]. http://saisrivatsa.com/Files/aes.pdf
[27] Office of China National Committee for Chinese Proficiency Test. The Grammar Section in A Grade Syllabus for HSK[M]. Higher Education Press, 1996.
[28] Phandi P, Chai K M A, Ng H T. Flexible domain adaptation for automated essay scoring using correlated linear regression[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 431-439.
[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
[30] Song W, Zhang K, Fu R, et al. Multi-Stage Pre-training for Automated Chinese Essay Scoring[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6723-6733.

APPENDIX

Table A1. Feature Set used in Regression AES Model

Level	Features
character	Total character count Square root of total character count Forth root of total character count Count of unique characters The ratio of number of unique characters to the total number of words in the essay Count of Grade A characters Count of Grade B characters Count of Grade C characters Count of Grade D characters Count of characters above Grade A to D Count of misuse characters
word	Total word count Unique word count Count of Grade A words Count of Grade B words Count of Grade C words Count of Grade D words Count of misuse words Average count of words in sentences
sentence	Total sentence count Total clause count The ratio of number of clauses to the total number of sentences The average sentence length The total count of sentence errors

AUTHORS

Lulu Dong (1996-)
Postgraduate student
HuiNing, Gansu, China;
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing

Lin Li (1980-),
Associate Professor
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing; Natural Language Generation

HongChao Ma (1979-)
PhD
HuaiNan, Anhui, China;
College of Intensive Studies
Beijing Language and Culture
University;
Research field: Language
Testing and Teaching
Evaluation.

YeLing Liang (1997-)
Postgraduate student
PingYao, Shanxi, China;
College of Computer of Qinghai Normal University;
Research field: Natural Language Processing

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

ONTOLOGY-BASED QUESTION ANSWERING SYSTEM FOR AN ACADEMIC DOMAIN

García-Robledo Gabriela A, Reyes-Ortiz José A,
González-Beltrán Beatriz A and Bravo Maricela

Departamento de Sistemas, Universidad Autónoma Metropolitana,
Unidad Azcapotzalco, Mexico

ABSTRACT

The development of question answering (QA) systems involves methods and techniques from the areas of Information Extraction (EI), Natural Language Processing (NLP), and sometimes speech recognition. A user interface that involves all these tasks requires deep development to improve the interaction between a user and a device. This paper describes a Spanish QA system for an academic domain through a multi-platform user interface. The system uses a voice query to be transformed into text. The semi-structured query is converted into SQWRL language to extract a system of ontologies from an academic domain using patterns. The answer of the ontologies is placed in templates classified according to the type of question. Finally, the answer is transformed into a voice. A method for experimentation is presented focusing on the questions asked in voice and their respective answers by experts from the academic domain in a set of 258 questions, obtaining a 92% accuracy.

KEYWORDS

Question answering systems, Speech recognition, Ontologies, SQWRL patterns, Information Extraction, Natural language processing.

1. INTRODUCTION

Teaching, research, and diffusion activities take place in an academic environment. Facilities are available, such as computer laboratories, libraries, classrooms, and auditoriums to realize these activities. In addition, actors participate in the activities to execute tasks. Administrative staff, professors, and students are some involved actors.

Ontologies allow to store and depict information about activities or events. Ontologies structure the information semantically to facilitate queries. However, the query about an event information and its related aspects (people, time, and physical spaces) involves a user request. Due to the large amount of information generated in the academic domain, this query process requires computational processing and a transformation into formal languages.

The answer extraction process from a voice query is a challenge today. The main objective of this paper is to extract an answer from an academic system of ontologies. For this reason, the whole process begins with a voice question from a user transformed into text and then represented in a semi-structured query using the procedure developed in [1]; the answer is placed in defined templates depending on the type of answer extracted; finally, the text with the final answer is converted from text to speech. This process involves a speech-to-text and text-to-speech translator, enriching a semi-structured query, and mapping entities to ontology elements. Also,

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 13-25, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111902

the identification of structural patterns (succession of structured elements frequently identified in different questions), the SQWRL execution query to get a specific response about events, physical spaces, and involved actors (people) that belong to the academic domain, and the generation of an answer understandable and friendly for the user is connected. The system of ontologies uses modules to consider people, time, and physical spaces; It is possible to answer questions such as who? where? and when? To make the query in voice format input and output, it uses Web Speech API in Java to run it in a multi-platform environment.

This paper aims to present a classification of templates based on different types of answers obtained from an information extraction method with an approach based on SQWRL patterns. As well, describe the development of a multiplatform user interface that has as input a voice query from a user from the academic domain, a transformation to an unstructured text, a natural language processing, extraction, and search of information in a system of ontologies, a representation of the answer in templates and voice output to the initial user.

The rest of the paper is organized as follows. Section 2 is focused on the most important and recent advances in the research areas of this paper. In Section 3, the proposed method for transforming the voice query into an SQWRL query is to extract information from the ontologies system and get an answer, which is returned to the user in the same voice format. The evaluation process is presented in Section 4, including experiments with 258 questions considering speech-to-text and text-to-speech. Finally, Section 5 presents the conclusions and future work.

2. RELATED WORK

In this section, a review of the related work is discussed and presented.

A natural language interface is presented in [2] called FREYA that uses SPARQL searches on ontologies; it executes natural language request and generates an answer in the form of a graph, it offers the option of choosing between several options in case of ambiguity in the question to train the system. In [3], Ginseng uses static grammar rules that provide English structures and phrases. In [4] QuestIO System is developed, which is an interface without a predefined vocabulary. It uses a dictionary to identify classes, relations, instances, and property values based on ontology knowledge.

The system presented in [5] warns drivers about critical situations they may encounter on their journey; it is implemented with front and back cameras in their vehicle. The driver interacts with the system using only the voice. The driver observes a critical situation, activates the voice system, and describes the situation. Then, the system recognizes the voice of a driver, identifies the keywords, and transfers the information to a central system. The authors use ontologies that allow structuring semantic annotations by type of accident with the help of keywords and using indicators (green, yellow, and red) to know the priority with which it should be attended. In [6], a support system is developed for warehouse employees handling chemical materials using their hands and sight; it is implemented in devices so that with the help of voice recognition, they can record the tests without diverting attention from the sample. The voice records are made in an ontology, and the location is used to recognize what type of articles are used based on the knowledge of the ontologies.

A personal assistant is presented in [7] based on natural language processing, carried out through parse trees. The system receives a sentence from the user through voice; if it was understood satisfactorily, it notifies and then executes the query to the database, finally returning the answer the same way as the request. In [8], the system can simultaneously perform tasks to support a user by voice. Among the tasks it can handle are searching for a document or editing it, handling

emails, and an activity schedule. The system is developed in multi-agent system architecture. In [9], a personal door assistant is implemented using voice; it allows a user to help manipulate the entrance to his office by acting as an intermediary between the visitor and the office owner. In [10], a portable personal assistant is patented to manipulate the contacts user, activities, and appointment scheduling using voice recognition. The system implements a voice user identifier to restrict access to personal data.

Furthermore, [11] developed a system to control several devices through applications within a house, known as Smart Home; the system uses voice to control devices and ontologies represent the information. The main aim is to execute specific tasks within the environment. The user requests and the dialogue system search the ontology to send the satisfactory operation to the Smart Home system. Later the extracted information is returned to the dialogue system, and it returns a report to the user as an answer. In [12], a system is presented that turns a space into an intelligent space for the service of the elderly to help them control electronic devices and different services using only the voice without the need for prior knowledge of the operation of the device.

There are works focused on people with limited vision, such as [13], whose search engine helps people with limited vision interact with a computer system in the same way as others, using only the voice. The visually impaired person communicates to the device with short sentences about what he/she requests on the internet, the system converts the voice sentences into text, and then it carries out the query to the ontologies that help to search the internet, the system displays the results of the URLs as the answer.

Finally, [14] also proposes a voice-manipulated search engine for blind people. This platform receives the voice with the user request, and it is converted into text. Then, the ontologies are consulted based on extracted keywords and thus support the search on the internet. The system returns the URLs that satisfy the requirement; they are converted to voice to answer the person who initiated the request. The authors comment that the platform helps in human-computer interaction with blind people. In [15], the QA system has input a query in natural language and an ontology, and as a response, it is extracted from the semantic markup of the compatible ontology. They propose a cascade architecture, where the query is translated into a set of triples compatible with the ontology. Furthermore, the natural language interface of [16] for the deduction of information stored in ontologies receives queries through voice in English. They developed a knowledge generator module (MGC) that answers user queries through the interface module (MI).

With the review of the work related to the disciplines involved in this paper, it is possible to observe the importance of carrying out an adequate transformation from a voice query into a structured representation and then using templates classification has been a transformation from text to speech.

3. QA SYSTEM PROPOSED

For the development of the QA System, a method was designed and implemented that transforms a query issued in natural language through a voice interface into a query in SWRL language; the answer is returned in the same input format. This process includes a voice input and output, a mapping of ontological entities, identifying structural patterns, executing an SQWRL query to obtain an answer from a system of ontologies, and classifying templates that allow better interaction between the user and the device, as in Figure 1 shows the architecture of the proposed Question-Answering System.

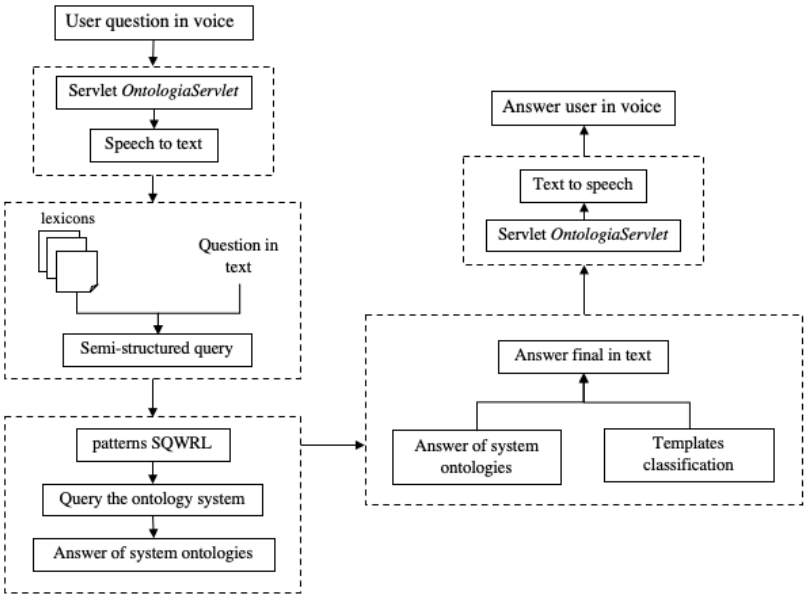


Figure 1. Proposed Question-Answering System Architecture

3.1. Speech to text

The user makes the question by voice. The Web Speech API in Java was used for speech recognition. JavaScript is used for the speech recognizer to transform the speech into text. When speech recognition is performed, it refers to a servlet named *OntologiaServlet*. The servlet names all the classes in charge of natural language processing and searching in the ontologies. The code is available at: https://github.com/UamAISII/AppVoz.git.

The answer generated in step 3 is used for the speech recognizer to transform the text to speech and is sent to the query interface.

3.2. Query processing

The query voice transformed into text is converted into a semi-structured query; this is created following the methodology reported in [1] by identifying entities and components of academic domain queries with a precision of 96%. The method in [1] is made from lexicons of events, physical spaces, and people in an academic environment, available at: https://github.com/UamAISII/AppVoz.git. They analyzed a system of ontologies named Intelligent Environment and developed in [17], which contains Person, Event, Sensor Network, Physical Space, and Time. An information retrieval structure and an ontological tuple are created with a particular structure, preserving unstructured elements [1]. This semi-structured query is improved in [18] to consider elements necessary to answer the request.

The semi-structured query in [18] has various elements, such as the type of question and the set of relevant entities that the request contains. The relevant entities are those that contain the information necessary to respond to the query. This set of information is represented in a 5-tuple of elements. An answer from the system of ontologies is presented using enrichment to the structure of [1] and the identification of 3 structural patterns in SQWRL to query the ontology system. The method used in [18] has a precision of 96.8%, and this is available at: https://github.com/UamAISII/AppVoz.git.

3.3. Answer generated

The answer received from the SQWRL query is placed in templates with some elements of the ontological mapping for each type of question developed in [18], the representation shown in Equation 1.

$$[O,S,P,TP,C,P2,TP2] \quad (1)$$

Where:

- O is the ontology or ontologies to be consulted.
- S is the subject extracted from the question.
- P is the property of the subject, for example, if it´s the name of a person, *hasName* is placed or if it´s the name of an auditorium, *hasNamePhysicalSpace* is placed. These properties correspond to the model of the ontology system.
- TP is the *DataProperty* type or *ObjectProperty* type.
- C is the class of the subject contained in the queried ontology and contains the information to answer, for example, *Professor, Building*.
- P2 is the property that is asked, for example, if we want to know the student ID, this data will contain *hasStudentID*.
- TP2 is the type of P2 and can be *DataProperty* or *ObjectProperty*.

In a *Person* question, the question is asked by name, employer ID (economic number), student ID, email, or category. If the question is for the category of a professor or student ID, the template is formed as follow:

La + P2 + *de* + S + *es* + ?res
(The + P2 + of + S + is + ?res)

For example, the question is: *¿Cuál es la categoría de Pérez Pedro?* (What is the category of Pérez Pedro?), the answer generated is: *La categoría de Pérez Pedro es profesor titular B* (The category of Pérez Pedro is full professor B).

In the case of asking for the name, economic number, or email of a person, the answer is represented:

El + P2 *de* + S + *es* + ?res
(The + S + P2 is + ?res)

For example, the question is: *¿Cuál es el correo de Pérez Pedro?* (What is the e-mail of Pérez Pedro?). The answer generated is: *El correo de Pérez Pedro es pperez@azc.uam.mx* (The e-mail of Pérez Pedro is pperez@azc.uam.mx).

For questions of the physical space type, it must be analysed if Equation 1 contains the *ObjectProperty constrainsPhysicalSpace* in the position of TP2; if so, the answer is generated as follows:

S + P2 + ?res

For example, the question is: *¿Qué oficinas contiene el edificio C?* (What offices does build C contain?). The answer generated is: *Edificio C contiene librería, auditorio Incalli Ixcahuicopa, recursos humanos, lenguas extranjeras, covi, cajas* (Building C contains a library, Incalli Ixcahuicopa auditorium, human resources, languages, covi, cashier).

In case of not contain the *ObjectProperty containsPhysicalSpace*, the answer is:

S + *está* + P2 + ?res
(S + is + P2 + ?res)

For example, the question is: *¿Dónde está el auditorio F001?* (Where is auditorium F001?), the answer generated is: *Auditorio F001 está ubicado en edificio F planta baja* (Auditorium F001 is in building F, ground floor).

The event type question type has three different templates; these depend on the answers received in the SQWRL queries. If the question is for the description of an event, the answer is:

S + *trata de* + ?res
(S + tries to + ?res)

For example, the question is: *De qué trata el taller paisajismo 101* (What is landscaping workshop 101 about?) the answer generated is: *Taller paisajismo 101 trata de reflexionar sobre la importancia de la construcción y diseño; para ello es importante conocer los recursos lineamientos y criterios básicos que se manejan en espacios exteriores* (Landscaping workshop 101 tries to reflect on the importance of construction and design; for this it is important to know the resources, guidelines and basic criteria that are handed in outdoor spaces).

If the question is to obtain the event date from its name, the answer is formed with the variables ?res and ?res2. The answer is:

S + *se realizará el* + ?res1 + *de* ?res2
(S + will be held on + ?res1 + ?res2)

For example, the question is: *¿Cuándo es el taller de prevención de violencia contra las mujeres?* (When is the workshop for the prevention of violence against women?), the answer generated is: *Taller de prevención de violencia contra las mujeres se realizará el 17 de Agosto* (Workshop for the prevention of violence against women will be held on August 17).

Also, it asks for the name of the event from a date, the answer is:

El nombre del evento es + ?res
(The name of the event is + ?res)

For example, the question is: *¿Cuál es el nombre del congreso del 18 de junio?* (What is the name of the congress on June 18?). The answer generated is: *El nombre del evento es congreso de cohetes hidropropulsados* (The name of the event is the hydro-powered rocket congress).

The questions of the *event-person* are the participants of an event from the date; in this case the answer is:

En el evento del + S + *participa* + ?res
(In the event on + S + participate + ?res)

For example, the question is: *¿Quién participa en el congreso del 15 de enero?* (Who participates in the congress on January 25th?). The answer is: *En el evento del 15 de enero participa Pérez Pedro, Gómez María* (In the event on January 15th participate Perez Pedro, Gómez María).

In the case of asking for the participants of an event with the name:

En + S + *participa* + ?res
(In the + S + participate + ?res)

For example, the question is: *¿Quién participa en el seminario de física?* (Who participates in the physics seminar?). The answer is: *En seminario de física participa Pérez Pedro, Gómez María.* (In the physics seminar participate Pérez Pedro, Gómez María).

The questions of the *physical space–event* type are classified in two. If the question is about the place where an event takes from the date, the answer is:

El evento del + S + *se realizará en* + ?res
(The + S + will be held in + ?res)

For example, the question is: *¿Dónde es el taller del 18 abril?* (Where is the workshop April 18?). The answer is: *El evento del 18 de abril se realizará en laboratorio E306* (The April 18 event will be held in laboratory E306).

Also, it asks for the name of the event; in this case the answer is represented:

S + *se realizará en* + ?res
(S + will be held in + ?res)

For example, the question is: *Dime dónde es el taller de prevención de violencia contra las mujeres* (Tell me where is the workshop for the prevention of violence against women). The answer is: *Taller de prevención de violencia contra las mujeres se realizará en laboratorio Steve Jobs* (Workshop for the prevention of violence against women will be held in the Steve Jobs laboratory).

The questions of the *physical space – person* type, can be by the person who is assigned to a workplace, the answer is:

El + S + *está asignado a* + ?res
(S + is assigned to + ?res)

For example, the question is: *¿A quién le pertenece el cubículo H261?* (Who owns cubicle H261?), the answer is: *El cubículo H261 está asignado a Pérez Pedro* (Cubicle H261 is assigned to Pérez Pedro).

Finally, when the question is about a person workplace based on their name or economic number, the answer is:

S + *se encuentra en* + ?res
(S + is in + ?res)

For example, the question is: *¿Cuál es el cubículo del profesor Pérez Pedro?* (What is the cubicle of Professor Perez Pedro?), the answer is: *Pérez Pedro se encuentra en cubículo H286* (Pérez Pedro is in cubicle H286).

3.4. Query interface

JSP (*JavaServer* Pages) technology connects the graphical query interface and the developed natural language processing.

The query interface is designed to help the user; when the query interface is opened, the user observes a window with a green button, as in Figure 2.

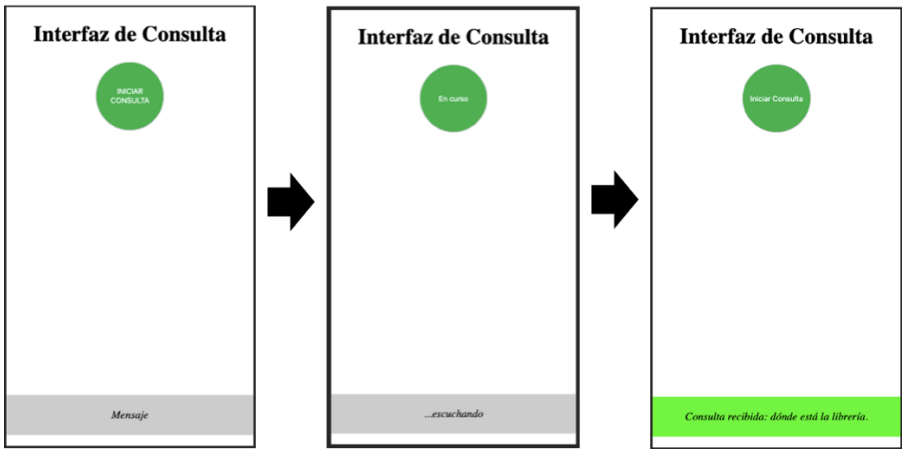


Figure 2. Query interface design

The user presses the *Iniciar consulta* (Start query) button and can begin to ask by voice; on the interface see a legend …*Escuchando* (... listening) is shown to indicate that you can start talking and the button changes to *En curso* (In progress). As a final part, the friendly voice of the system mentions the answer in Spanish that it obtains from the ontology system and shows the query at the bottom, so the user can know it was recognized in the query interface. A use case is a bellow:

Use case "Juan wants to know where the library is":

When Juan presses the *Iniciar consulta* (start query) button, the system goes to the listening state and waits for the query. Juan asks by voice the question *¿Dónde está la librería?* (Where is the library?). When the system listens for a long pause, it terminates the listening state and goes into an in-progress state. During this period, the system performs query processing and the answer generated.

If the system recognizes the query, it is displayed at the bottom as *consulta recibida: dónde esta la librería* (Query received: Where is the library). Moreover, the interface terminates the progress state and into a speaking state where the answer is heard through voice: *La librería está en el edificio C planta baja* (The library is in building C, ground floor). If it is not recognized, the answer is: *Por el momento, no puedo responder* (At the moment, I cannot answer).

4. EXPERIMENTATION AND RESULTS

For evaluation purposes, a method for experimentation is assessed and executed. This method utilizes data that is extracted from an academic domain ontology [16]. This ontology contains 855 individuals, 91 classes, 10736 axioms, 36 object relations, and 51 properties or datatype relations. A total of 258 questions were defined by experts (research faculty members, students, and assistants), the language of the questions generated in Spanish. Seventy experts participated in defining the questions, specifically professors, assistants, and students. From these questions, a corpus was integrated. Experimentation method and corpus are available at: https://github.com/UamAISII/AppVoz.git.

The QA system developed can be used in mobile devices with two operating systems, such as Android or iOS. The precision of each question was evaluated manually by experts.

The speech recognizer has a margin of error due to different situations, such as ambient noise, user diction, or unknown words. It was considered part of the evaluation of whether the voice recognizer identified the complete question or if an absence of any word influences the system to answer correctly (partially identified question). For example, if the user asks ¿Dónde se encuentra la oficina del profesor Pedro Pérez? (Where is the office of Professor Pedro Pérez?), and the recognizer only identifies se encuentra oficina del profesor Pedro Pérez (is the office of Professor Pedro Pérez).

Table 1 shows speech recognition evaluation for the *Person* question-type. For this issue, 11 questions were asked: 79 questions were fully recognized, and 32 questions were wrong. In addition, out of 32 questions that were partially recognized, 27 correct responses were obtained, and of the 79 fully recognized questions, 74 successful responses were obtained.

Table 1. Person question-type speech recognition performance evaluation.

Question voice recognition	**Recognized questions**	**Correct answer**	**Wrong answer**
Fully identified	79	74	5
Partially identified	32	27	5
TOTAL	**111**	**101**	**10**

Nineteen questions of the *PhysicalSpace* question type were applied. In Table 2, it is observed that 14 questions were fully recognized, and five questions were not. Of the 14 fully recognized questions, 92.86% correct answers were obtained, and of the five partially recognized questions, all questions were answered successfully.

Table 2. PhysicalSpace question-type speech recognition performance evaluation.

Question voice recognition	**Recognized questions**	**Correct answer**	**Wrong answer**
Fully identified	14	13	1
Partially identified	5	5	0
TOTAL	**19**	**18**	**1**

Table 3 shows the speech recognition evaluation of the *Event* question type. For this type, 35 questions were completely identified, and five questions were not, of the 40 questions that were asked for this type. Of the five partially recognized questions, all were effectively obtained; and of the 35 fully identified questions, 88.57% of successful responses were obtained.

Table 3. Event question-type speech recognition performance evaluation.

Question voice recognition	Recognized questions	Correct answer	Wrong answer
Fully identified	35	31	4
Partially identified	5	5	0
TOTAL	**40**	**36**	**4**

For the *Person-PhysicalSpace* composite question-type, eight questions were asked, five were fully recognized, and three were partially recognized. All questions were answered successfully. Table 4 shows the evaluation information for this type of question.

Table 4. Person-PhysicalSpace question-type speech recognition performance evaluation.

Question voice recognition	Recognized questions	Correct answer	Wrong answer
Fully identified	5	5	0
Partially identified	3	3	0
TOTAL	**8**	**8**	**0**

In the *Person-Event* composite question-type, 40 questions were asked, 34 were fully identified, and six partially. All fully identified questions were successfully answered, and partially identified questions obtained the 66.66% of correct answers. The results for this type of question are detailed in Table 5.

Table 5. Person-Event question-type speech recognition performance evaluation.

Question voice recognition	Recognized questions	Correct answer	Wrong answer
Fully identified	34	34	0
Partially identified	6	4	2
TOTAL	**40**	**38**	**2**

Table 6 shows the speech recognition evaluation for the *Event-PhysicalSpace* question-type. Thirty-six questions were fully recognized for this type, and four parts of the 40 questions were asked. Of the four identified questions, three were successfully obtained, and 94.44% of fully recognized questions obtained successful answers.

Table 6. Person-Physical space question-type speech recognition performance evaluation.

Question voice recognition	Recognized questions	Correct answer	Wrong answer
Fully identified	36	34	2
Partially identified	4	3	1
TOTAL	**40**	**37**	**3**

The graphic in Figure 3 shows the results of the evaluation. One hundred ninety-one questions were fully identified, and 47 questions were partially identified, both with a correct answer. For questions with an incorrect answer, 12 were fully recognized questions, and eight were partially identified.

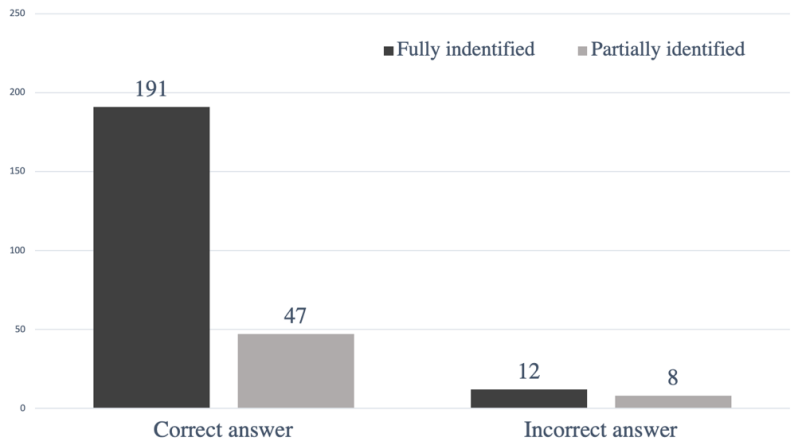


Figure 3. Global question-answer performance evaluation

The results of the global performance evaluation were obtained with the well-known metric in the Information Extraction area *accuracy*, shown in equation 2.

$$P = (Relevant\ Results) / (total\ questions) \quad (2)$$

Where:

- Relevant results. Correct answers are fully identified, and correct answers are partially identified by the speech recognizer.
- Total questions. Total questions asked by experts (research faculty members, students, and assistants).

The results obtained were encouraging, achieving 92% of *accuracy*.

5. CONCLUSIONS

This paper has presented a QA system in the Mexican Spanish language within an academic environment. In addition, an ontology-based approach is described as an information representation model to perform queries in SQWRL from a voice query using structural patterns.

In the tests, promising results were obtained. The architecture of the QA system involves the following contributions: (a) developing a multi-platform voice query interface made with Web Speech API. (b) the construction of an ontological 5-tuple from natural language processing techniques in Spanish; (c) the identification and implementation of three structural patterns with syntax in the formal SQWRL language for the query in a system of ontologies of the academic domain; (d) the adaptation of an information extraction method based on lexicons in Spanish to a query architecture; (e) generating answer friendly to the user with templates selected according to the question type.

The results obtained in section 4 show that the QA system for voice is promising, having a more significant number of correct answers fully and partially identified. These results suggest an advance considering that they were carried out with the noise of an academic environment and the diction of different people who belong to the academic community. The question answering system showed that the number of correct answers is not affected by the partially recognized questions.

As future work, (1) it is intended to check the functionality of the query interface in the Spanish language of Spain, to identify if the accent influences speech recognition; (2) enrich the semantic model in real-time; (3) adapt the query interface so that it can answer in real-time the location of the participants; (4) implementation of the QA system in another domain, for example, health;(5) Perform an approach of extraction and identification of answers to texts with deep learning algorithms and replace ontologies; (6) Carry out a measurement of the quality of the QA system with the mathematical model presented in [19] that predicts the degree of satisfaction of the interested parties (Q) that constitute quality characteristics of the software.

REFERENCES

[1] García, G., Reyes, J., González & B., Priego, (2019) "Extracción de información a partir de preguntas en español basada en lexicones", Academia Journals, Vol. 11, No. 8, pp1054-1059.

[2] Damljanovic, Danica, Agatonovic, Milan & Cunningham, Hamish, (2010) "Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction". The Semantic Web: Research and Applications 7th Extended Semantic Web Conference, Vol. 6088, No. 1, pp106-120.

[3] Bernstein, Abraham, Kaufmann, Esther, Kaiser, Christian & Kiefer, Christoph, (2006) "Ginseng: A guided input natural language search engine for querying ontologies", Jena User Conference.

[4] Tablan, Valentin, Damljanovic, Danica & Bontcheva, Kalina, (2008) "A Natural Language Query Interface to Structured Information", Proceedings of The Semantic Web: Research and Applications, Proceedings of the 5th European Semantic Web Conference, Vol. 5021, No. 1, pp361-375.

[5] Sosunova, Inna, Zaslavsky, Arkady, Theodoros, Anagnostopoulos, Alexey, Medvedev, Sergey, Khoruzhnikov & Cladimir, Grudinin, (2015) "Ontology-based voice annotation of data streams in vehicles", Internet of Things, Smart Spaces, and Next Generation Networks and Systems, 15th International Conference, Vol. 9247, No. 1, pp152-162.

[6] Kopsa, Jirí, Míkovec, Zdenek & Slavik, Pavel, (2005) "Ontology driven voice-based interaction in mobile environment". International Conference and Research Center for Computer Science.

[7] Rawassizadeh, Reza, Dobbins, Chelsea, Nourizadeh, Manouchehr, Ghamchili, Zahra & Pazzani, Michael, (2017). "A natural language query interface for searching personal information on smartwatches". Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference, pp679-684

[8] Paraiso, Emerson & Barthès, Jean-Paul, (2006). "An intelligent speech interface for personal assistants in R&D projects". Expert Systems with Applications, Vol. 31, No. 4, pp673-683.

[9] Yan, Hao & Selker, Ted, (2000) "Context-aware office assistant". Proceedings of the 5th international conference on Intelligent user interfaces, pp276-279.

[10] Tsiao, J. C. S., Chao, D. Y. & Tong, P. P., (2007) "Natural-Language Voice-Activated Personal Assistant" U.S. Patent No. 7,216,080. Washington, DC: U.S. Patent and Trademark Office.

[11] Huang, Cheng-Chi., Liu, Alan & Zhou, Pei-Chuan, (2015) "Using Ontology Reasoning in Building a Simple and Effective Dialog System for a Smart Home System", Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference, pp1508-1513.

[12] Portet, Francois, Vacher, Michel, Golanski, Caroline & Meillon, Brigitte, (2012) "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects", Personal and Ubiquitous Computing, Vol. 17, No. 1, pp127-144.

[13] Bukhari, Ahmad & Kim, Yong-Gi, (2012) "Ontology-assisted automatic precise information extractor for visually impaired inhabitants". Artificial Intelligence Review, Vol. 38, No. 1, pp9-24.

[14] Karthik, N., Ashwini, M. & Anitha, K., (2014) "Voice Enabled Ontology Based Search Engine on Semantic Web For Blind", International Journal of Computer Science & Engineering Technology, Vol. 5, No. 04, pp341-344.

[15] Lopez, Vanessa, Pasin, Michele & Motta, Enrico, (2005) "Aqualog: An ontology-portable question answering system for the semantic web", European Semantic Web: Research and Applications Second European Semantic Web Conference, Vol. 3532, No. 1, pp546-562.

[16] Solís, Alejandro, Florencia, Rogelio, Acosta, Carlos & López, Francisco, (2018) "Interfaz de lenguaje natural para deducción de información almacenada en ontologías". Research in Computing Science, Vol. 147, No. 6.

[17] Padilla, Josué, (2019) “Detección y representación de eventos en un ambiente académico inteligente”, Master's Thesis, Universidad Autónoma Metropolitana, Unidad Azcapotzalco.
[18] Gabriela A. García-Robledo, Beatriz A. González-Beltrán, José A. Reyes-Ortiz & Maricela Bravo, (2020) “Extracción de respuestas a partir de ontologías utilizando patrones estructurales en SQWRL”, Research in Computing Science, Vol. 149, No. 8, pp571-585.
[19] Gheisari, M., Panwar, D., Tomar, P., Harsh, H., Zhang, X., Solanki, A., Nayyar, A. & Alzubi, J., (2019). “An optimization model for software quality prediction with case study analysis using MATLAB”. IEEE Access, Vol. 7, pp 85124-85138.

AUTHORS

Gabriela A. García-Robledo is a partial-time researcher-professor at the Universidad Autónoma Metropolitana. She received a master’s in computer science from Universidad Autónoma Metropolitana in 2020. Her current research interests include natural language processing, speech recognition and information extraction.

Jose A. Reyes-Ortiz received a Ph.D. in Computer Science from National Centre for Research and Technological Development, Cuernavaca, Mexico in 2013. Currently, he is a full-time researcher-professor at Universidad Autónoma Metropolitana. His current research interests include natural language processing, pattern recognition from text, and computational linguistics.

Beatriz A. González-Beltrán is a professor of Computer Science at the Universidad Autónoma Metropolitana. Her work focuses on artificial intelligence in human computer interaction. Beatriz completed her Ph.D. in the Institute of Engineering at the Univ. Grenoble Alpes, France.

Maricela Bravo obtained the degree of Doctor of Computer Science at the National Centre for Technological Research and Development (CENIDET) in 2006. She currently works as a full-time professor at the Autonomous Metropolitan University. Her research areas include knowledge representation, logical reasoning, distributed systems based on REST architectures, and intelligent systems for healthcare applications.

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

DEVELOPMENT OF ADMINISTRATION PROFESSIONAL COMPETENCES IN BRAZILIAN PUBLIC UNIVERSITIES: A MULTICASE STUDY IN FLORIANÓPOLIS

Ednaldo de Souza Vilela[1], Filipe José Dias[2] and Marcos B. L. Dalmau[2]

[1]Municipal University Center of São José, São José, Brasil
[2]Federal University of Santa Catarina, Florianópolis, Brasil

ABSTRACT

The article aims to investigate how the development of competences applied to the professional formation of the egress administrator of public municipal higher education institutions in the Florianópolis region occurs under perspective of teachers and coordinators of the bachelor's degree in administration course. For this, a qualitative and documentary research was carried out, using a structured questionnaire applied to 20 people as a data collection instrument, including 2 course coordinators and 18 professors from the studied institutions who teach the subjects whose contents are related to professional formation from the administrator. The results show that the new national curriculum guidelines encourage the development of competences. In this context, despite the effort to comply with such devices, there is some misalignment between the teaching plans and the pedagogical project of the course. Difficulties in implementing formation based on competence and lack of institutional stimuli are also perceived.

KEYWORDS

Development of competences, Higher Education, Formation, Administration Course.

1. INTRODUCTION

The Brazilian model of higher education, regulated by the Ministry of Education of Brazil, classifies the Higher Education Institutions - IES observing the Law 9.394/96 regarding the academic organization in two types: university institutions (universities, specialized universities and university centers) and non-university institutions (Federal Technological Education Centers (CEFETs in portuguese) and Technological Education Centers (CETs in portuguese), integrated faculties, isolated faculties and higher education institutes).

Also according to the aforementioned law, regarding the administrative organization, HEIs are classified as public educational institutions when created, incorporated, maintained and administered by the Federal, State or Municipal government or private institutions when they are maintained by individuals or legal entities of private law and for profit or not (community, confessional and philanthropic).

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 27-44, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111903

After the promulgation of the Brazilian Constitutional Charter on October 5, 1988, the municipal HEIs that charged monthly fees up to that year kept this right, however, the municipal HEIs that were created after that, lost this right.

In the Florianópolis region, there are two free municipal public HEIs. The Municipal University Center of São José – USJ, the first municipal university center, public and free in Brazil, and in the municipality of Palhoça there is the Municipal College of Palhoça – FMP.

Both have in common the offer of the undergraduate course in Administration, with a percentage between 70% and 80% of vacancies in their selection processes (vestibular) for its citizens, considering the municipality's need for professional qualification, aiming to meet the demands of the local market and regional, economic, social and sustainable development.

Degree courses in Administration as a business school, positioned as one of the most offered courses in the Brazilian context, in order to serve the labor market, undergo reformulations in order to adapt them to the desired professional profiles for the graduates of the course, guiding adhering to principles such as curriculum flexibility and dynamism, emphasis on general education and competences development.

Since formation aimed at developing competences requires changes in the teaching-learning process, even though there are normative standards that guide the construction of the curricula of administration courses throughout the Brazilian territory, through the National Curriculum Guidelines (DCN in portuguese), it was raised the following question: How is the development of competences in the professional formation of administrators in public municipal HEIs in the Florianópolis region under perspective of teachers and course coordinators?

In order to answer this question, the general objective of this study was to investigate how the development of competences applied to the professional formation of administrators in public municipal HEIs in the Florianópolis region, under perspective of teachers and course coordinators, takes place. the specific objectives are: (i) it seeks to ascertain whether the competences described in the National Curriculum Guidelines of the Administration course are included in the Courses Pedagogical Projects of the studied HEIs; (ii) analyze subject programs to verify if they express which competences are developed (professional subjects); (iii) identify practices used to develop competences in the professional formation of administrators (professionalizing disciplines) and (iv) propose strategies (Identify through research carried out good practices that should be stimulated/multiplied) to meet the needs of competences development.

It is important to highlight that this study is delimited to the analysis from the perspective of municipal HEIs in the Florianópolis region, responsible for the formation of intellectual capital in these cities in recent years, whose offer of the aforementioned on-site course was also due to the demand of the region, which claimed qualified human resources to compose the companies' workforce. The study is justified by providing the professors and coordinators of the researched Business Administration course with information about the profile of the administrator that companies in Florianópolis expect, so that they can self-assess and provide subsidies for these studied HEIs to assess whether there is a need to renew its Pedagogical Policies.

2. THEORETICAL FOUNDATION

2.1. Competence

Defining competence is not a simple task, due to its use in different areas and under different perspectives.

First of all, we know and understand that there are two terms and meanings about the main subject: competence and competency. The first one refers to an individual's capacity to perform and fulfill job responsibilities. In another point of view, competency focuses on an individual's actual performance in a particular situation.

For P. Perrenoud (1999), competence is the ability to articulate a set of schemes, thus going beyond knowledge, enabling knowledge to be mobilized in the situation, at the right time and with discernment.

To A. Zabala (2010),

The use of the term "competence" is a consequence of the need to overcome a teaching that, in most cases, was reduced to a memorizing learning of knowledge, a fact that implies difficulty for this knowledge can be applied in real life.

The competency-based formation model has, at its foundations, the orientation of formation for the development of competences that are replicable in the work environment (F. Vargas et al, 2001).

Conforming to the studies by E.P. Rossoni (2013), there are three currents that deal with the topic: American, English and French (A.S. Godoy.Et.Al., 2009; and R.C. Guimarães, 2009) and that when carrying out a literature review A.S. Godoy et al. (2009, p.267) observed that among the authors used a common point about the notion of competence is its derivation based on the “set of knowledge, skills and attitudes expected from people, the so-called in portuguese language as “CHA” .

M.T.L. Fleury (2002) defines competence as the junction between theoretical knowledge (knowledge – knowing) the ability (task – knowing how to do) and being an attitude (attitude – knowing how to be).

The combinations of these three resources (knowledge, skills and attitudes) or competence dimensions, applied together at work, can explain the competence of a person in conformity to G. Le Boterf (1999).

The dimensions of competence were clarified by S.T. Bergue (2014, p. 263), since, according to him, knowledge is "those conceptual or technical elements that a person has or needs to have to perform a certain activity", the competences allude to the "ability to transformation of knowledge into action" and cites as examples of skills communication, analytical capacity, flexibility, and persuasiveness.

Finally, the aforementioned author clarifies that attitudes are related to personality attributes and personal and professional posture, which reveal the "impulse of the agent for action" and examples of these are: ethical values, transparency, frankness, courtesy, cordiality, respect , among others.

Other authors refer to attitude as social and affective aspects related to work, or a person's positive or negative reaction, their predisposition to adopt a specific action, or even a feeling, an emotion or a degree of acceptance or rejection of the person in relation to others, objects or situations (T. Durand, 2000, R.M. Gagné et.al 1988).

For M.R. Banov (2012), knowledge is the domain, the clear and correct understanding of the information in the area of expertise; skill in applying the technique, the ability to put knowledge into practice; and attitude when acting, decision making when required (dimensions of competence).

2.2. Higher education: degree in Administration

In historical terms, according to E.P. Rossini (2013), it was in 1952 that the first administration course was created, and the most important educational institution in Brazilian administration was founded in 1944, Fundação Getúlio Vargas - FGV, originating from national politics developmentist of President Vargas.

However, for L. Siqueira and S. Nunes (2011), the expansion in the number of administration courses is accentuated by the opening to foreign capital, based on the regulation of the profession and education, reinforced by the implementation of the government's development policy of President Juscelino Kubitscheck (1956-1960).

On September 9, 1965, through Law No. 4,769, the profession in conforming to E.P. Rossini (2013), restricting access to the professional market to holders of titles issued by the university system, and, shortly thereafter, in 1966 and later in 1993, it had its minimum curricula approved resulting in the proposal of Curriculum Guidelines for the course of Administration in 1988.

From the point of view of L. Siqueira and S. Nunes (2011), the expansion of undergraduate courses in administration occurred, mainly, in isolated colleges of the private sector, with greater concentration in the southeast and south regions of the country, whose opening was still advantageous as it does not require high investment for its implementation.

Based to data from MEC/Inep//Daes-Enade/2018, 1,765 administration courses were evaluated in the ENADE 2018 exam, of which 515 courses (29.2%) were offered at universities, colleges presented 931 courses (52, 7%), the university centers offered 292 (16.5%) and the Cefet/Ifet, in turn, offered 27 courses, corresponding to 1.5% of the total courses.

Some authors such as A. Nicolini (2003), the accelerated expansionism of undergraduate courses in administration would result in the massification of teaching and will require an innovative pedagogy so that it does not stop at stagnation, whose alternative would be for formation based on competences whose proposal discusses pedagogical methods, which oppose the “mass” formation movement in educational institutions.

Since the legislation pertaining to the teaching of Administration remained unchanged until 1993, it was with the promulgation of the new Law of Guidelines and Bases of National Education, LDB n. 9,394/96 (BRASIL, 1996), of December 20, 1996, which reopened a new debate on education, whose highlight was the insertion of the notion of competences in higher education in Brazil (L. Siqueira and S. Nunes, 2011); .

In conformity to the above authors, the Basic Guidelines Law n. 9,394/96 (LDB) allowed for the flexibility of course curricula, as long as they were linked to their curricular guidelines.

2.3. National curriculum guidelines - (DCNs)

In the context of formation based on competences, this debate was intensified after Resolution n. 4/2005 of July 13, 2005, established by the National Council of Education, together with the Chamber of Higher Education, instituting the National Curriculum Guidelines for the Graduate Course in Administration, as a guideline for the preparation of curricula in institutions of University education.

Pursuant to Resolution No. 04/2005:

1st The course's Pedagogical Project - PPC, in addition to the clear conception of the undergraduate course in Administration, with its peculiarities, its full curriculum and its operationalization, will cover, without prejudice to others, the following structural elements: (...) Art. 5 Undergraduate courses in Administration must include, in their pedagogical projects and in their curricular organization, contents that reveal interrelationships with the national and international reality, according to a historical and contextualized perspective of its applicability within organizations and the environment through the use of innovative technologies and that meet the following interconnected fields of forming: I - Basic Forming Contents: related with anthropological, sociological, philosophical, psychological, ethical-professional, political, behavioral, economic and accounting studies, as well as those related to communication and information technologies and legal sciences; II - Professional Forming Contents: related to specific areas, involving theories of administration and organizations and the administration of human resources, market and marketing, materials, production and logistics, financial and budgeting, information systems, strategic planning and services; III - Contents of Quantitative Studies and their Technologies: covering operational research, game theory, mathematical and statistical models and application of technologies that contribute to the definition and use of strategies and procedures inherent to administration; and IV - Complementary Forming Contents: optional transversal and interdisciplinary studies to enrich the trainee's profile.

This Resolution n. 4/2005 also deliberates on which competences and abilities the course should form, as described below:

I - recognize and define problems, equate solutions, think strategically, introduce changes in the production process, act preventively, transfer and generalize knowledge and exercise, in different degrees of complexity, the decision-making process; II - develop expression and communication compatible with professional practice, including in negotiation processes and in interpersonal or intergroup communications; III - reflect and act critically on the sphere of production, understanding its position and function in the production structure under its control and management; IV - develop logical, critical and analytical reasoning to operate with values and mathematical formulations present in formal and causal relationships between productive, administrative and control phenomena, as well as expressing themselves critically and creatively in the face of different organizational and social contexts; V - have initiative, creativity, determination, political and administrative will, willingness to learn, openness to change and awareness of the quality and ethical implications of their professional practice; VI - develop the ability to transfer knowledge of daily life and experience to the work environment and their field of professional activity, in different organizational models, proving to be an adaptable professional; VII - develop the capacity to prepare, implement and consolidate projects in organizations; VIII - develop the capacity to carry out consultancy in management and administration, administrative, managerial, organizational, strategic and operational opinions and expertise. (BRASIL, 2005, p. 2).

For A. Nicolini (2003, p. 54), “desirable competences to the administrator, when they are not innate, must be developed throughout the course, a development that assumes the student as the subject of their own forming process” (A. NICOLINI, 2003, p. 54).

Based in a point of view of A.B. Silva et.al. (2007), in the professional forming process, students can help to understand the dynamics between learning and competences development if they act as an active participant in the process, which is influenced by various contexts – academic, professional and personal (social).

However, there is a new debate around higher education in Administration, since after a broad study carried out in mid-2020 by the commission of the National Education Council of the Ministry of Education (CNE/MEC), with the participation of the Federal Administration Council (CFA) and the National Association of Undergraduate Courses in Administration (Angrad), resulted in a consensus on the need to update the new National Curriculum Guidelines (DCNs) for the Bachelor's Degree in Administration course. Such changes, according to this committee, are a reflection of the development of the technological field and the market, and among the highlights of the new DCN are forming through competences and mandatory professional practice.

3. METHODOLOGICAL ASPECTS

This research, in order to achieve the objectives of the study, adopted an applied research, which based in the understanding of F. Appolinário (2011, p. 146), is carried out with the aim of “solving problems or concrete and immediate needs”. often, these problems emerge from the professional context and can be suggested by the institution so that the researcher can solve a problem-situation.

As for the objectives, the research is documentary, because, conforming to N. Tumelero (2019), the documentary sources can be used through tables, formal documents, photos, meeting minutes, and various reports, so that in the future, the desired results can be obtained.

In this research, the PDI of the studied HEIs were used as documentary sources, as well as the Course Pedagogical Projects of their Administration courses. Regarding the approach to the problem, this is a qualitative study that is appropriate when “the researcher seeks to establish the meaning of a phenomenon from the points of view of the participants” (J.W. Creswell, 2010, p. 42). This research refers to two case studies in undergraduate courses in Administration and two municipal HEIs, free and public, located in the Florianópolis region, Santa Catarina.

In the point of view of M. Ludke and M.E.D.A. André (1986, p. 17), the case study, it is always well delimited, and its outlines must be clearly defined in the course of the study. The case may be similar to others, but it is at the same time distinct, as it has its own unique interest. The interest, therefore, focuses on what is unique, what is particular, even if certain similarities with other cases or situations later become evident.

The data collection instrument was structured with the objective of evaluating the perception of the coordinators of the two Administration courses of the researched HEIs and of their professors of the professional forming content disciplines, whose phenomenon (the development of competences applied to the professional formation of the administrator under perspective of teachers and coordinators) was empirically investigated in its real context. Questionnaires were used as data collection instruments that were sent to coordinators and teachers of subjects whose contents are related to the professional forming of undergraduate courses in Administration of the

municipal HEIs of Palhoça and São José, totaling 20 teachers to whom the questionnaire was sent .

From the selected sample, responses were obtained from 2 coordinators and 14 professors, being 9 (nine) professors from the USJ and 5 (five) professors from the FMP, as shown in the following Figure 1:

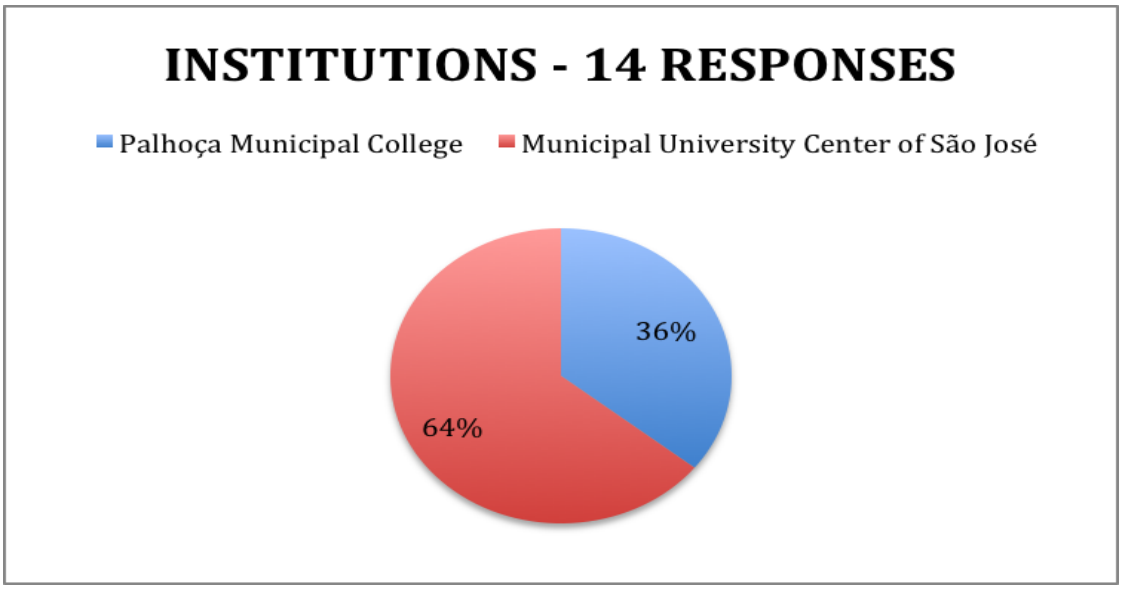


Figure 1 - Respondent Origin

In the analysis of this article, the respondents are identified as follows: C1 the coordinator of the (USJ) and C2 the coordinator of the (FMP). Among the professors, 9 (nine) are from USJ and identified themselves citing at this time also their discipline, such as: U1 (General Theory of Administration), U2 (Project Management), U3 (Material Management), U4 (Finance Business), U5 (Management Information System), U6 (Production Management), U7 (Marketing), U8 (Strategic Business Management), U9 (Human Resources) and 5 (five) of the FMP: F1 (People Management) , F2 (Financial Administration), F3 (General Theory of Administration), F4 (Market Research) and F5 (Market Administration). In the following topic, the analyzed results are discussed.

4. DOUBLE CASE STUDY: MUNICIPAL HEIS

4.1. Municipal University Center of São José – USJ

The Municipal University Center of São José (USJ) was created by Municipal Law 4279 of May 15, 2005, and is the first public and free Municipal University Center in Brazil to offer the opportunity of access to free higher education with a view to entrepreneurship. and innovate the practice of municipal development with social responsibility, according to its Institutional Development Plan -PDI (2015-2019). Also based of this document, the USJ is located in the municipality of São José in Santa Catarina, and was approved by the State Council of Education on July 12, 2005, by unanimous vote, to approve the request for its accreditation granted by a favorable opinion. the operation, for two years, of the following courses: Administration, Pedagogy and Accounting.

The USJ instituted the affirmative action public policy, with social vacancies, with 70% (seventy percent) of the total vacancies offered every six months for students who attended educational institutions maintained by the government (state, federal and municipal), being the 30% of the remaining places available to any Brazilian or foreign citizen who wants an opportunity in higher education. Then, in conformity to the PDI (2015-2019), it obtained the renewal of its accreditation for a period of five years, based on Resolution no. 27, and Opinion no. 122, of the State Council of Education, signed on May 22, 2007, and later on September 24 and 25, 2015, the institution was evaluated by a new Commission constituted by the State Council of Education, and was successful in continuing its operation, through a final concept of 3.76.

The Graduate Course in Administration at USJ is authorized to offer 80 (eighty) annual places at night, in a credit system, being divided into 40 (forty) places for admission in the first semester and 40 (forty) places for admission in the second semester, through a unified selective process (vestibular) of the Acafe System and by transfer and return, according to its Pedagogical Course Project - PPC (2020).

Based on this PPC (2020) the curriculum matrix of the Administration Course is composed of 42 theoretical subjects (2652 hours) and 1 practical subject related to the curricular internship (80 hours), organized conforming to a prerequisite logic, which must be followed during the development of the students' study, prepared accordingly with the syllabuses of the subjects of the current curricular structure. After telephone contact with the current coordinator of the Administration Course at USJ, we obtained agreement to carry out our research with her professors.

4.2. The Municipal College of Palhoça - FMP

The Municipal College of Palhoça - FMP is an autarchy created by Municipal Law No. 2.182, of October 25, 2005, an entity that is part of the indirect public administration of the Municipality of Palhoça with legal personality under public law, being its sponsor the Municipality of Palhoça, whose Statute and General Regulations of the FMP were elaborated in accordance with the requirements of Law 9394/96 (Law of Guidelines and Bases of National Education), which was updated and is published as Decree n. 1489/2013, based of to its Institutional Development Plan (2019-2023).

Considering to this same document, its accreditation was given by an act of the State Council of Education, which accredited the FMP by Opinion No. 056 and Resolution No. 016, of April 4, 2006, and which through Law 2386 of June 21 2006, with 80% of the vacancies in its courses being reserved for students from public secondary schools residing in the municipality, equalizing the opportunities for admission to higher education and the other 20 c/o are available to any Brazilian or foreign citizen who wishes an opportunity in higher education. In 2010, the FMP Administration course was evaluated and recognized with a 4.02 concept by the State Council of Education of Santa Catarina.

In the terms of the Pedagogical Project of the Course - PPC (2020) of Graduation in Administration of The Municipal College of Palhoça - FMP, which is based on the Curriculum Guidelines of the Graduation in Administration Course, Resolution No. 4, of July 13, 2005, it obtained authorization based on Resolution No. 016 and Opinion No. 056 approved on 04/04/2006, recognized again for a period of 04 (four) years, based on Resolution No. 101, Opinion No. 293 of December 7, 2010, approved by Decree State n.1930, published by DOE n. 19,726 of 12.18.2013.

Also in consideration of this document, it obtained authorization for the operation of the expansion from 100 (one hundred) vacancies to 200 annual vacancies for the undergraduate course in Administration at the Municipal College of Palhoça. Opinion 213 and Resolution CEE 112 of 08.28.2012, and Renewal of Recognition of the Bachelor's Degree in Administration course. Opinion no. 189 and Resolution n. 089 of 12.08.2015. In the aforementioned Course Pedagogical Project, it is stated that the Curriculum Matrix 2016.1 of the Graduate Course in Administration in force was approved by CONFAP Resolution 010/2016. after contact by telephone with the current coordinator of the Administration Course at FMP, we obtained agreement to carry out our research with their professors.

5. RESULTS

Coordinators and professors asked about "What do you understand by competence?" the following answers shown in the table below were obtained.

Table 1. Understanding of competence according to respondents

(C1) Competence is the set of knowledge, skills and attitudes (cha)	U1 Set of knowledge, skills and attitudes necessary to carry out activities effectively, efficiently and effectively.
(C2) Consists of the individual's abilities to learn and develop technical and behavioral skills that can be put into practice.	U2 Set of knowledge, skills and attitudes
F1 These are the knowledge and/or abilities of an individual, perceived by other people. It configures a person's behavior as it is put into practice.	U3 Competence is a person's ability to use their know-how and skills to do a good job.
F2 Make the theory I teach useful in practice. That students can apply. This is my competence. Proper Didactics	U4 Competence is a person's ability to use their know-how and skills to do a good job.
F3 Develop skills	U5 It is an organizational requirement that encompasses employees' knowledge, skills and attitudes. Some theorists define attitudes as behavior.
F5 Knowledge and characteristics to carry out an activity.	U6 The attributions that someone has to exercise something...
	U7 Competence is the professional's baggage of knowledge, skills and behaviors.
	U8 Concept involving people's knowledge, skills and competences
	U9 Based in M.T.L. Fleury (2002), knowing how to act is responsible and recognized, which implies mobilizing, integrating, transferring knowledge, resources, skills that add economic value to the organization and social value to the individual. For those who are starting the administration course, we teach that competence is the application of C.H.A. (Knowledge = knowing; Skills = knowing how to do; Attitudes = knowing how to be). Without delivery, there is no competence.

From the results to this question, it can be seen that half of the respondents understand the concept of competence referring to the set of knowledge, skills and attitudes. Of the other respondents, some are linked to the idea of behavior.

Coordinators and professors asked about "What competences are provided to graduates of the Administration Course?" varied responses were presented as shown in the following table.

Table 2. Competences of Graduates according to interviewees

C1 I think they must have knowledge, skills for the profession and necessary attitudes	U1 Technical and behavioral knowledge, inter and intrapersonal skills and ethical attitudes.
C2 Preparation of organizational diagnosis for solving complex problems; Ability to understand the social, political, economic and cultural environment in which it operates and make decisions that ensure the sustainability of the organization and the environment; Carry out planning in different areas of administration considering multiple scenarios; Ability to undertake a business; Ability to work in a team, developing interpersonal relationships, leadership and conflict management; Development of critical sense, logical reasoning and interpretation	U2 Manage projects
F1 Carry out the diagnosis to find opportunities and solve problems in organizations; Develop planning in different areas of administration; Ability to undertake in business; Ability to work with people and manage conflicts; etc.	U3 Entrepreneurial and business management skills.
F2 Amplified vision, action planning on what was measured.	U4 Entrepreneurial and business management skills.
F3 Being able to manage a business	U5 Theoretical knowledge of different areas of administration (people management, marketing, finance, logistics), skills with administrative tools (using specific systems and solving problems) and attitude (interpersonal relationships, teamwork, holistic view)
F4 Critical thinking, sense of urgency, organization	U6 Learning, technical and also living skills, communication
F5 Practical activities around marketing.	U7 Theoretical and empirical knowledge in the areas of administration, management, technical skills as well as group and individual performance behaviors.
	U8 Leadership, business vision, good communication, etc.
	U9 Scientific knowledge; ability to analyze complex problems and take action. decision; career development; critical approach; leadership development; team work; ability to use their technical-scientific knowledge to generate process improvements (innovation); development of written communication and presentation; analyze scenarios; among others.

In this aspect, it can be inferred that most of the answers obtained are incomplete and/or are not in accordance with the competences defined in the National Curriculum Guidelines. This may

demonstrate a lack of a broad vision of the role of a particular discipline in the development of competences in professional formation.

When the coordinators were asked about the relationships between competences described in Table 2 and the Teaching and Course Pedagogical Project programs, they replied that they are fully or partially described in the programs and/or teaching plans of the subjects and that such competences are in accordance with the competences provided for in the Course Pedagogical Project. However, the competences listed in the PPC are partially developed in the Course's disciplines.

The answers in figure 2 were compared with the teaching plans of the disciplines and it was found that the competences mentioned are included in the plans as objectives of the disciplines, but many do not use the term “competence”.

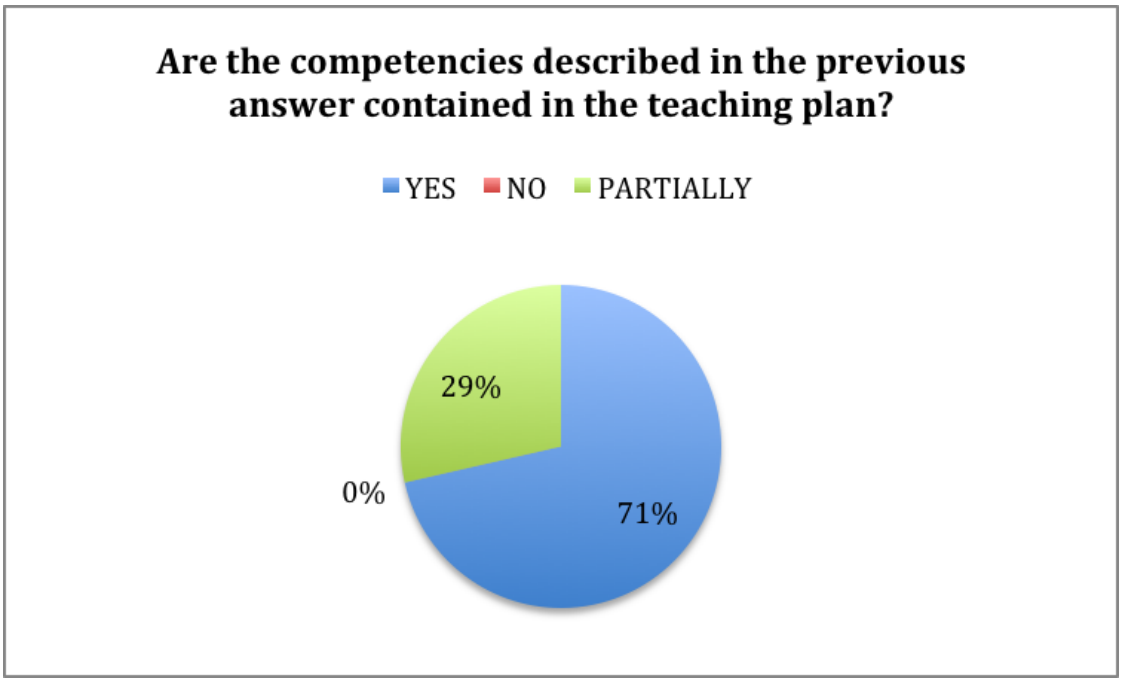


Figure 2 - Competences and Teaching Plans in the view of coordinators and the teachers

Analyzing the responses of the course coordinators as shown in figures 2, 3 and 4, it is possible to see that there is no direct association between the competences defined in the Course Pedagogical Project and the competences developed in the practices of the disciplines.

In this aspect, it appears from the As a result, due to the lack of articulation between the subjects, some competences defined in the Course Pedagogical Project are not developed during the course, which can cause some fragility in professional formation.

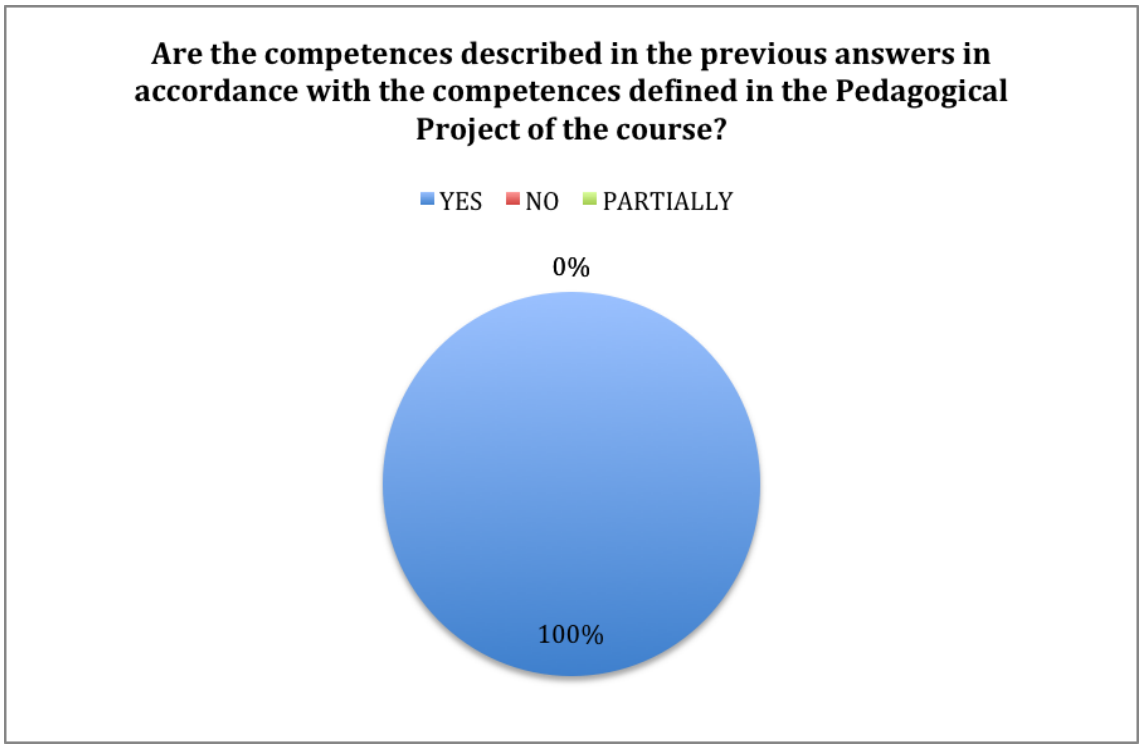


Figure 3. Competences and Course Pedagogical Project

On the other hand, the importance of developing competences is perceived when teachers are asked about the level of contribution of their discipline to the development of competence in the general education of the student.

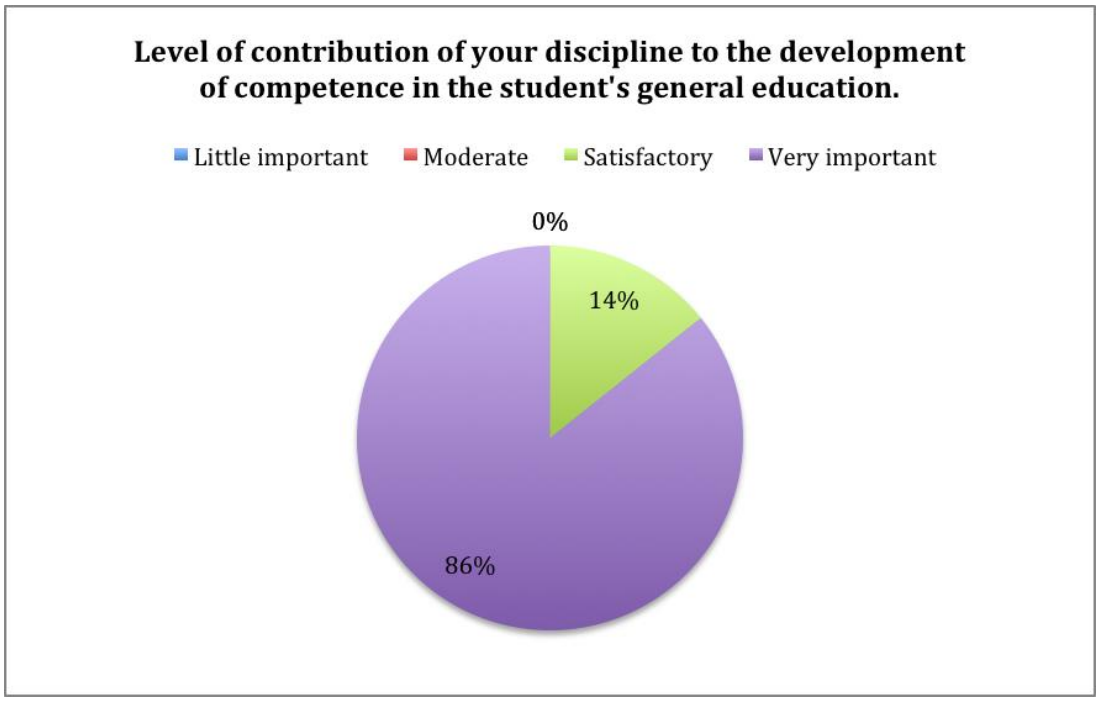


Figure 4. Competences, Course Pedagogical Project and disciplines

Regarding the professors, asked about the strategies or tools used to develop competences in the discipline they are responsible for, the professors and coordinators responded as shown in table 3 below.

Table 3 - Strategies or tools used to develop competences?

C1 Projects, etc.	U1 Team practice of theoretical issues worked on.
C2 The subjects must comprise theoretical, practical and field classes in their workloads, which promote the student's protagonism in the process of developing competences	U2 Cases, Mind Maps, Canvas
F1 Practical activity for the preparation of the TD&E program, following the four stages of planning, described in a didactic way for the student to develop; Dialogue presentation of the themes of the syllabus, seeking to awaken interest in the student, reconciling his day to day; Case studies for critical analysis and problem solving; Personal development, based on studies and research that promote the identification of individual strengths and how they can be balanced for teamwork and conflict management within organizations. Flipped classroom. The topic is made available and the student needs to research to solve the problems proposed by the teacher.	U3 Case study, application of mathematical and statistical models, debates.
F2 Real examples.	U4 Analysis of statements of real companies, investment simulation, interpretation of financial information.
F3 Discussion and criticism of theories	U5 Lectures on theoretical, historical and systems structure study. Practical case study with systems development for a chosen company, preferably a small one that has everything to develop. Reflective analysis by academics of the development of each company presented by them.
F4 Group work, Field research	U6 Practical group work, with research and practical exercises, videos...
F5 relationship marketing is the main one.	U7 Content explanatory lessons in conjunction with practical lessons. Every theory envisions the applied practice in which the student performs the marketing function or analyzes a practical case.
	U8 Case Studies and Concept Discussion
	U9 Teamwork; case studies; Presentations; Preparation of reviews from technical scientific readings; Debates; etc.

From the answers presented in Table 3, the teaching effort can be seen, but there is still some difficulty in defining and applying practices aimed at formation by competence.

The same can be seen in the respondents' answers when they are asked to share experiences of strategies or tools used for the development of competences (table 4), but in this aspect there is a greater variety of instruments.

Table 4. Share experiences of strategies or tools used to develop competences.

C1 360 degree rating.	U1 Strong weight in the practices of reading, analyzing and interpreting scientific articles with further debate and conclusions on the topic addressed.
C2 The projects will be developed by each teacher in his/her discipline that makes up the extension workload, and should serve organizations and the Palhocense community.	U2 design canvas
F1 Exercises with the steps of a TD&E program; Case Study; Quizz; Researches; Mapping of organizational competences, using Dashboard in Excel	U3, U4 and U8 not answered
F2 BPs and DREs of SA.	U5 First, I make a theoretical overview so that they have a basis on the themes. Then we go to the practical part to check the organizational reality. In the practical part they end up having to perform with their own abilities, creativity and interactivity.
F3 Case Study Suggestions in Current Times	U6 I use Metimeter, Jamboard as course tools, as well as videos and games built by students with course content
F4 Google forms	U7 The strategy of putting product analytics into practice in every promotional compound. They usually participate when they look for a product in their own home to identify the logo and the loyalty relationship with the product and the packaging.
F5 I develop a marketing project that they learn to work in the marketing area.	U9 Students are given a topic to research and later prepare for a class debate. On the day of class, I divide the class into two groups: the first group will defend the theme and the second group will oppose the theme, even if the student has a different view than the one he will defend. This makes it possible to exercise a critical view of the problems

Another relevant aspect in competency formation is assessment. In this sense, the professors were asked: "How are the assessments carried out to verify the development of competences in the scope of the discipline?"

Table 5. Assessment information by competence

The assessments so far are not done by competences. But to meet the DCNs, it is intended to expand the assessment methods in the discipline, making it clear to students the competences that are being assessed in each teaching-learning strategy, with a scale of 0-10 for the competence being developed in that activity. Self-assessment and teacher assessment can be done.
Through the presentation of practical work
Practical presentation of a project
Individual and team assessment
Through a set of critical reviews and debates.
Discursive evaluations involving practical cases
Evaluation where the mastery of concepts, the capacity for practical analysis and financial management in a company, investment analysis is verified.
Financial reports and stock projection

Assessments are carried out during the application process in the organizational reality. There is a need for the student to be active and it is possible to measure their skills and attitudes precisely in this context. The transformation of the theoretical part into practice demonstrates the learning capacity and competences applied in the organization.
At this time of Covid 19 pandemic, in which we are remotely teaching assessments through exercises, case studies, assignments and tests
Through the project that is done in stages.

In the aspect of evaluation, it is noted among the answers presented in Table 5, that there is a greater variety of instruments used that aim to fulfill the task of verifying the achievement of objectives or competences in each discipline, but in this study it is not possible to define whether such instruments are sufficient and adequate for the objectives they set themselves.

And the last question was whether the institution encourages pedagogical practices aimed at formation through competences, whose answers are in figure 5:

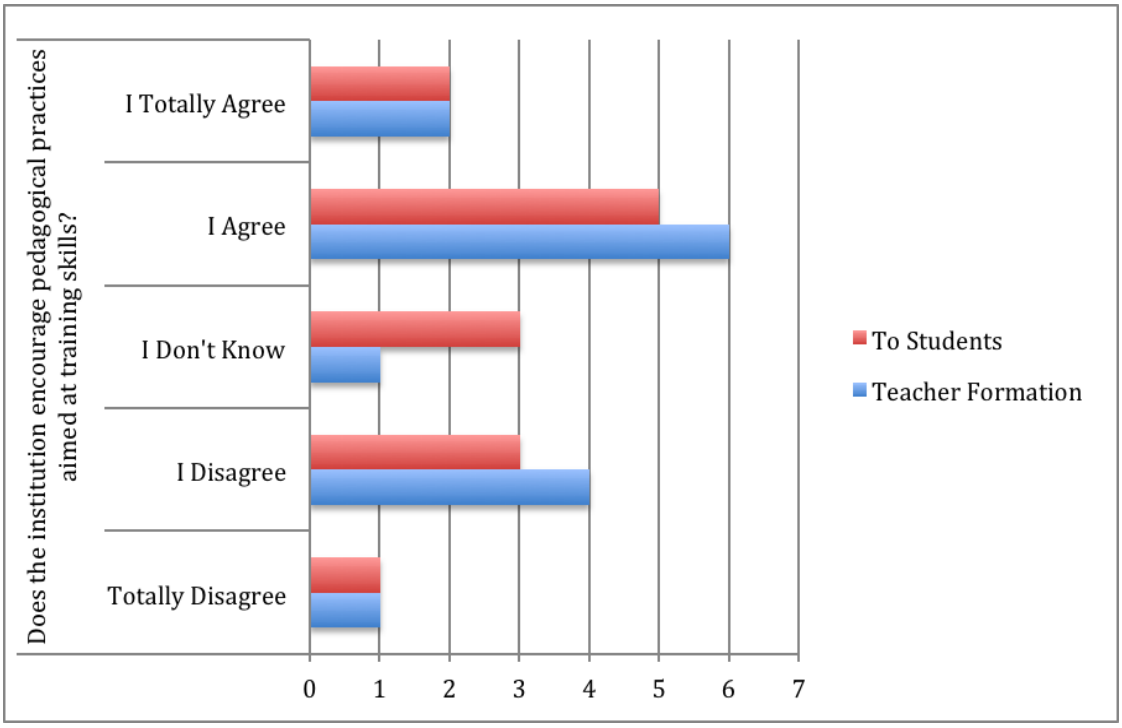


Figure 5. Stimulus Pedagogical Practices - Coordinators' and Teachers' Response

Among the responses obtained, there is a very sensitive aspect for formation based on competence in the education of students, which is the institutional incentive, both to offer support and formation to teachers and to promote a new teaching-learning format among students.

Realizes the need for the institution to promote alignment and constant formation for articulation between the competences defined in the DCN and PPC of the courses that must be developed in the subjects, and the definition, adjustments, or support for the development of strategies to promote formation based on competence.

The practice of integrated activities between the subjects and the encouragement of student practices to carry out other oriented activities outside the classroom, such as extension activities, can constitute alternatives to encourage interaction with real situations and help in the maturation

of competences necessary for the professional practice. However, to confirm this hypothesis, further observations and studies that can solidify this view are needed.

6. CONCLUSION

The main objective of this study was to verify how the development of competences applied to the professional formation of the egress administrator of the public municipal HEIs in the Florianópolis region takes place under perspective of professors and coordinators of the Bachelor's Degree in Administration course.

Therefore, a theoretical survey was carried out on the understanding of competence and the competences defined in the National Curriculum Guidelines of the Administration course and the PPCs of the Courses present in 2 municipal public institutions in Florianópolis region were analyzed.

In the analysis of the obtained results, a lack of articulation and/or association between the competences defined in the National Curriculum Guidelines, Course Pedagogical Project and those that are developed in the disciplines can be seen.

Analyzing the responses of the course coordinators, it is possible to see that there is no direct association between the competences defined in the PPC and the competences developed in the practices of the disciplines. In this aspect, it appears from the result that in the absence of articulation between the disciplines, some competences defined in the Course Pedagogical Project are not developed during the course, which can cause some fragility in professional formation.

Among the results obtained among teachers, there is an effort to develop competences in the subjects, but there is some lack of clarity in understanding the concept of competence. The difficulty is even more accentuated when it comes to the use of a strategy for the development of competences, lacking clarity about competence-based formation.

Realizes the need for the institution to promote alignment and constant formation for articulation between the competences defined in the PPC of the courses that must be developed, and the definition, adjustments, or support for the development of strategies to promote formation based on competence.

REFERENCES

[1] F. Appolinário, Dicionário de Metodologia Científica, [Scientific Methodology Dictionary] 2. Ed, Atlas, São Paulo, 2011.
[2] M.R. Banov, Recrutamento, seleção e competências. [Recruitment, selection and competences], 3. Ed, Atlas, São Paulo, 2012.
[3] S.T. Bergue, Gestão estratégica de pessoas no setor público. [Strategic people management in the public sector], Atlas, São Paulo, 2014.
[4] BRASIL. Lei no 9394, 20 de dezembro de 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm.
[5] BRASIL. Ministério da Educação. (2005). [Classificação das instituições de ensino superior]. Disponível em: http://www.mec.org.br.
[6] J.W. Creswell, Projeto de pesquisa: métodos qualitativo, quantitativo e misto. [Research project: qualitative, quantitative and mixed methods] 3. Ed, Artmed, Porto Alegre, 2010.
[7] M.A. Cunha, A expansão do ensino de Administração em Curitiba e região metropolitana no período de 1997 a 2002, [The expansion of Business Administration education in Curitiba and the metropolitan region from 1997 to 2002], Seminário em Administração, 9, FEA – USP, 2006.

[8] T. Durand, L'alchimie de la compétence, [The alchemy of competence] Revue Française de Gestion, n. 127, 2000, p. 84 102.
[9] M.T.L. Fleury, As pessoas na organização. [The people in the organization] Gente, São Paulo, 2002.
[10] R.M. Gagné, L.J. Briggs, W.W. Wager, Principles of instructional design, Holt, Rinehart and Winston, Orlando, 1988.
[11] A.S. Godoy, C.S. Antonello, D.S. Bido and D. Silva, "O desenvolvimento das competências de alunos formados do curso de Administração: Um estudo de modelagem de equações estruturais." [The development of competences of graduated students of the Administration course: A study of structural equation modeling]. Revista de Administração - RAUSP, 44 (3), 2009, pp. 265-278.
[12] R.C. Guimarães, Representação social e formação da consciência crítica no curso de graduação em administração da EA-UFRGS, [Social representation and formation of critical awareness in the undergraduate course in administration at EA-UFRGS], doctoral dissertation, Porto Alegre, 2009.
[13] G. Le Boterf,, Competence et navigation professionnelle.[Competence and professional navigation] Éditions d´Organisation, Paris, 1999.
[14] M. Ludke, M.E.D.A de André, Pesquisa em educação: abordagens qualitativas, [Education research: qualitative approaches], Pedagógica e Universitária, São Paulo, 1986.
[15] Nicolini, "Qual será o futuro das Fábricas de Administradores?" [What will be the future of Administrator Factories?]. Revista de Administração de Empresas, São Paulo, v. 43, n. 2, 2003, p. 44-54.
[16] P. Perrenoud, Avaliação: da excelência à regulação das aprendizagens, [Assessment: from excellence to learning regulation], Editora Artes Médicas Sul, Porto Alegre, 1999.
[17] E.P. Rossoni, O desenvolvimento de competências na formação do administrador: um estudo na Universidade Federal de Rondônia, [The development of competences in the training of administrators: a study at the Federal University of Rondônia], doctoral tesis, Federal University of Rio Grande do Sul, Porto Alegre, 2013.
[18] A.B. Silva, A. Alberton, M.A. Verdinelli, As Competências Profissionais do Administrador e suas Implicações na Formação Acadêmica. [The Professional Competences of the Administrator and Their Implications for Academic Formation], I Encontro de Ensino e Pesquisa em Administração e Contabilidade, EnEPQ – Anpad, Recife, 2007.
[19] L. Siqueira and S. C. Nunes, "Um olhar sobre o Projeto Pedagógico e as práticas docentes baseados na Proposta de Formação por Competências." [A look at the Pedagogical Project and teaching practices based on the Skills Training Proposal]. Administração: Ensino e Pesquisa, vol. 12, núm. 3, julio-septiembre, 2011, pp. 415-445
[20] N. Tumelero, Pesquisa documental: conceito, exemplos e passa a passo, [Document research: concept, examples and step by step], https://blog.mettzer.com/pesquisa-documental/ (2019).
[21] F. Vargas, F. Casanova, L. Montanaro, El enfoque de competência laboral: manual de formación, [The focus on labor competence: training manual], Cinterfor/OIT, Montevideo, 2001.
[22] Zabala, Como aprender e ensinar competências. [How to learn and teach skills], Penso, Porto Alegre, 2010.

AUTHORS

Marcos B. L. Dalmau

Graduated in Administration from the Federal University of Santa Catarina (1999), Master's (2001) and Ph.D. (2003) in Production Engineering from the same institution. Full Professor at the Federal University of Santa Catarina. Professor of the Postgraduate Program in Administration (PPGA), at the level of Master's and Academic Doctorate and Professor of the Professional Masters in Administration University PPGAU/UFSC.
https://orcid.org/0000-0002-8620-1625
https://www.researchgate.net/ profile/Marcos_Dalmau

Ednaldo de Souza Vilela

Graduated in Accounting Sciences from the Federal University of Santa Catarina (1994) and Master in Production Engineering from the Federal University of Santa Catarina (2000). He is currently a professor - Centro Universitário Facvest, Centro Universitário Municipal de São José.

Filipe José Dias

Master in University Administration from the Federal University of Santa Catarina (UFSC) in 2019 and Bachelor of Law from the University of Vale do Itajaí (UNIVALI) in 2011. Doctoral Student in Administration - ppga/UFSC and scholarship holder of the UNIEDU/FUMDES Program. He is currently an Administrative Technician in Education at the Federal University of Santa Catarina.

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

AN INTELLIGENT DATA-DRIVEN ANALYTICS SYSTEM FOR OPERATION MANAGEMENT, BUDGETING, AND RESOURCE ALLOCATION USING MACHINE LEARNING AND DATA ANALYTICS

Dele Fei[1] and Yu Sun[2]

[1]St. Margaret's Episcopal School, 31641 La Novia Avenue,
San Juan Capistrano, CA 92675
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

This is a data science project for a manufacturing company in China [1]. The task was to forecast the likelihood that each product would need repair or service by a technician in order to forecast how often the products would need to be serviced after they were installed. That forecast could then be used to estimate the correct price for selling a product warranty [2]. The underlying forecast model in the R Programming language for all of the companies products is established. In addition, an interactive web app using R Shiny is developed so the business could see the forecast and recommended warranty price for each of their products and customer types [3]. The user can select a product and customer type and input the number of products and the web app displays charts and tables that show the probability of the product needing service over time, the forecasted costs of service, along with potential income and the recommended warranty price.

KEYWORDS

Operation Management, Machine Learning, Data Mining.

1. INTRODUCTION

This research was based on a manufacturing company's service department. The company's name is FastLink China. The products it produces are mostly industrial doors and dock levelers [4]. It is the top 1 in this criteria of business in China. Since it is still a developing company, there are hopes to make it better in minor parts. The goal for this project is to analyze the best price to maximize the profit in the service department. It is imperative to the company simply because the service department has the highest profit rate compared to other departments. However, the problem is that company owners have no idea how to set a price for extended warranty to help them gain the profit. Indeed, the research is based on the hope that this problem could be solved through data science skills. This can lead to a much larger profit in the company when they know the exact cost for the warranty and then determine profit based on different types of customers and industrial doors. Besides, it is a precious opportunity for me to learn about data application on businesses and to get familiar with how it runs and the way it works.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 45-56, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111904

There are a variety of tools or systems like fire base that have been used as a means for users to analyze their data for their personal use or business use. However, these existing tools are not that useful to me. Their implication and usage are too fundamental for the case since this is a research by setting up an analysis for a unique kind of data for a unique type of business [5]. It is not common that these existing tools seem limited. If the final result is provided by these tools, the accuracy will be doubtful and might have a huge influence since it is provided to a company that is related to money [6]. In this case, a more customized and sophisticated analytic tool is needed to provide a reliable and credible result to the company.

On the other hand, there is machine learning that is obviously workable in this situation. However, it is time consuming and too technical for a high school student to perfect the result also due to limitation of resources. Moreover, it is tough to explain the logic behind it to the people in the business company. Therefore, finding a better tool than normal, but simpler than machine learning is the goal [7].

Our goal is to forecast the price of warranties that will benefit the company. To this end, a survival curve is used inspired by insurance companies. Speaking of tools, the front end and back end both exists for the users. Front end is more HTML coding with apex charts and graphically displacement. For the back-end, tidy-verse and r are used for basic data cleaning. From ggplots and survival curve, the predicted percentage of breakdown for each month of each type of productcan be accessed [9]. Then, with basic calculations and taking the mean value of costs for appointments, an reliable outcome of recommended prices of warranty can be produced.

To the end of proving the result, admittedly, the most common ways will be to use train and test to see which has the best prediction. Also, r squared is a value that is usually considered by data scientists [8]. However, the situation is different from others since data that can be considered as correct to compare with the warranties does not exist. Moreover, there isn't enough data for us to train and split in some situations since some types of business and products only have a small amount of service history after group buys. Besides, the plot shows a strong curve just by plotting it through ggplots. Indeed, the research uses the approach to predict it through user surveys because the opportunity to let the manager class in the company determine if this data is applicable in real life situations or not exists. This is useful because all the people who took the survey will be familiar with all the products, companies, and potential users, which is the service department in this case.

Through looking at the official definition of user survey, a survey with 10 questions each with a rating from one to five is designed. After adding the rating for positive questions like "I think that I would like to use this system frequently", and minusing the rating for negative questions like "I found the system unnecessarily complex," the total score is doubled to normalize the total score like the official website asks. Indeed, getting the result over 68, which is the number to determine if the model is useful, proves that this development is effective.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges during the experiment and designing the sample; Section 3 focuses on the details of solutions corresponding to the challenges that was mentioned in Section 2; Section 4 presents the relevant details about the experiment, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Identifying the Problem and Approach

The most important challenge in this problem is identifying the problem and finding an approach to it. Beginning with a single conversation, the company owner complained about how it is tough for them to set a proper price for the extended warranty. Without it, he is concerned about how to plan for his companies' future since he does not have a predicted revenue. Instead of having the concern go in one ear and out the other, intuitively realize that data science could solve this problem, curiosity drives to investigate in how the price of the extended warranty with existing data can be forecasted [10]. Then based on the existing data, a unique approach emerges. After communication, the date of past services, cost and income of each service, and when the door starts to function exists in the database. Even though the realization of that number of dates is a vital variable to the prediction helps to get on the right path, going through linear models, logistic models, machine learning, and neutral networks, the correct approach is still ambiguous to this question. While the question lingers in mind, a collaborator who works in a data science company help on deciding a survival curve is usually used for the predictions for insurance companies that has the input of time elapsed for each product. Then, the connection between the average cost and benefit for each product will make the final step toward the final result.

2.2. Setting Price

One challenge in the problem is how to set a price for the extended warranty that is acceptable to the customers and profitable for the company. The company expresses their concern that they are only making guesses for the price that should be set for the warranty because they have no idea on the probability of the chance the specific type of product is going to break down. This problem is vital to the company's income because competing companies set a cheap price for their product, but an expensive price on their extended warranty [11]. To maintain the market, the company needs to set a price that is able to comfort the customers. In addition, they have never calculated the mean value of cost of labor and parts for fixing. Indeed, this model helps to solve this problem by investigating the number of breakdowns in an order for a period of time to get the percentage and uses the average cost to determine a price for the future orders.

2.3. Data

Initially, jetlag and distance between China and the United States forms a communication issue that makes the process of getting the data hard. Eventually, repeating the process of waiting and asking for more data, final result is slowly approached as the research develops. However, challenges come with opportunities. Since the workers of the company have not practiced standardized training on entering the data, not only the format of data is usually a mess to deal with, the emptiness also becomes a hot potato of the research. Needing to select a boundaries of data that eliminates the outliers and empty data without too much influence on the result is important. Also, there is not much data from the company so it is imperative to try to keep the amount of data [12]. Through identifying and excluding the empty data, the data for the initial prediction is cleaned. However, there are still some services that produce a negative profit for the company, which does not make sense. In this case, the company itself needs to find the correct input in order to solve this problem successfully.

3. SOLUTION

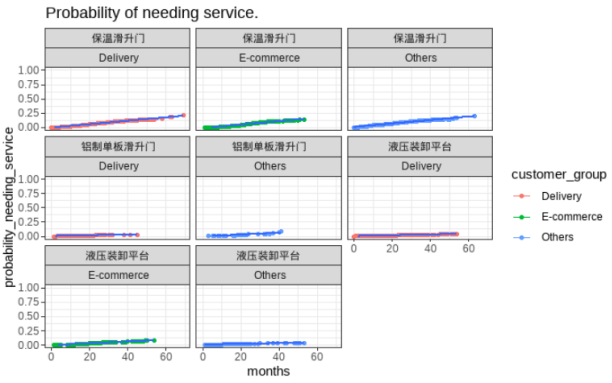


Figure 1. Probability of needing service

The scope of this project is to address the problems of … To break down the challenge into smaller programmable questions, these questions were generated:

What is the final outcome/indices that I'm trying to compute? What is necessary for me to compute such indices?

1. maintenance: (% of product need maintenance, and cost per maintenance [whether through average or median, and why did you choose it])
2. replacement: same as above

In general, the recommended price of extended warranty that can make the company predict the probability of breakdowns is the goal of this research. With the data of the door starting functioning date, needing service date, a survival curve is used like other insurance companies mostly use. Since the goal is to output the recommended warranty price, getting the center value of past historical services to know what the predicting cost and profit is required by combining two data. Eventually, the outliers are eliminated and just take the mean since the data is not heavily skewed. Indeed, the profit percentage is an input for the user for what they are looking for.

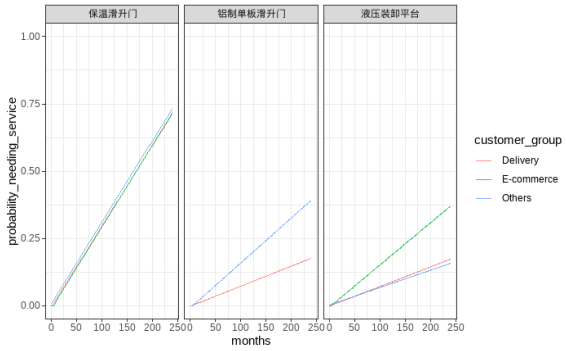


Figure 2. Breakdown percentage

Going into more detail step by step, graph of breakdown percentage is established using existing data to predict the trend. The x-axis represents the months, and the y-axis represents the percentage of this door needing service. The graph is divided into three products. Each of them is showing their major type of customer, which is normally the one that has enough data to predict

the trend. Since the business has the most customer groups in delivery and E-commerce, both types have the sufficient amount of data to make predictions. From the figure 2, the data has a strong correlation. It is great for future predictions.

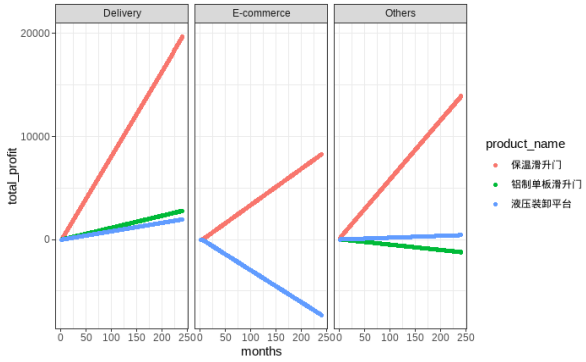


Figure 3. Profit

After getting the average cost and profit from the past data, these data with the percentage earlier are implemented to see the total profit that the business could gain after months for specific products in different businesses. This would be useful for the business since they could have a visualized understanding of its profit in the service department, which can be used for adjusting their price or percentage of profit based on their past data. It is obvious that there is some error in the existing data since there are some products that are bringing negative profits. This error might be due to the incorrect input from workers or the incorrect price they set for the customers by the business.

product_name <chr>	customer_group <chr>	n_total_products <dbl>	appointments_per_year <dbl>
保温滑升门	Delivery	100	3.6003095
保温滑升门	E-commerce	100	3.6674700
保温滑升门	Others	100	3.6426980
铝制单板滑升门	Delivery	100	0.8922874
铝制单板滑升门	Others	100	2.0091848
液压装卸平台	Delivery	100	0.8735344
液压装卸平台	E-commerce	100	1.8829967
液压装卸平台	Others	100	0.7751315

Figure 4. Appointments of each group

break_even_warranty_price <dbl>	recommended_warranty_price <dbl>	warranty_profit <dbl>
1301.2025	1951.8038	650.6013
1151.5917	1727.3876	575.7959
1375.2239	2062.8359	687.6120
329.9657	494.9486	164.9829
528.5626	792.8439	264.2813
506.3778	759.5667	253.1889
1060.6087	1590.9131	530.3044
536.5768	804.8653	268.2884

Figure 5. Prices and profit

Another feature the research produces for the company is a table that allows users to input the number of products as n_total_products and the percentage of profit they want to make. Since the model of percentage of needing services and the average cost of each appointment for different products both existed, this table is able to produce a recommended warranty price taking the percentage of profit the user is hoping to make. Also, the last graph of the businesses existing profit can assist in their decision to adjust the percentage of profits based on different products.

The table is also divided into three types of products and different types of customers. This is useful to help the company to set the warranty price.

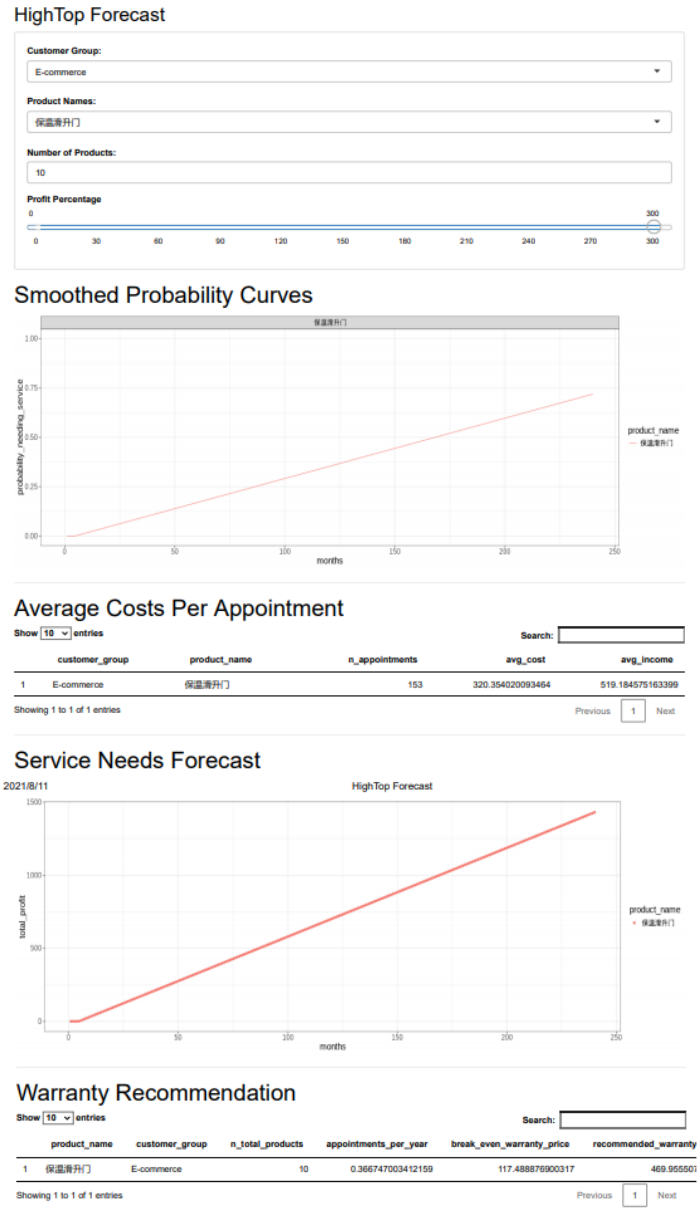


Figure 6. Warranty price

Besides pure coding, front-ends exists that directly help the users or business to understand the information they want. The input part allows business to adjust the customers' types, the products they are selling, the total number of products included in the warranty, and the percentage of profit. Then, the existing data will demonstrate a table and a chart. The chart displays the probability of needing services specifically that product and the type of customer. The table represents the average cost, profit, income of the specific product. Finally, it displays a graph of the probability of the breakdown, which is a prediction based on all the service history of this product. Last but not least, a warranty price recommendation is there for the user in the bottom table.

The choice of establishing a front-end is to better demonstrate data and encourage all the services workers to utilize it since it will be helpful eventually. If all of them are coding, the valuable information can only be accessible to a small number of workers, which decreases the efficiency of working.

In conclusion, this table fits the goal of displaying all the information that is helpful for the company to know their past data and the recommended future price for their future planning.

```
format_data <- function(){

  # this function takes the loaded data frames and cleans and prepares it so there is one row per product
  cat("Formatting data.", fill = T)
  service_data <- order_information %>%
    filter(product_name %in% c("保温滑升门",
                               "液压踏知平台",
                               "铝制单板滑升门")) %>%
    mutate(month_install = as.Date(paste(month(install_date), '1', year(install_date), sep = "/"), format = "%m/%d/%Y")) %>%
    group_by(customer_type, project_name, customer, product_name, month_install, warranty_length_zh) %>%
    summarize(
      install_date = min(install_date),
      product_quantity = sum(product_quantity)
      ) %>%
    select(-month_install) %>%
    left_join(
      google_translations %>%
        rename(warranty_length_zh = chinese,
               warranty = numeric) %>%
        select(-google_translate)
      ) %>%
    mutate(warranty_end_date = install_date %m+% years(warranty)) %>%
    uncount(product_quantity) %>%
    group_by(customer_type, project_name, customer, product_name, install_date) %>%
    mutate(item_number = 1:n()) %>%
    left_join(
      service_log %>%
        select(-project_zone, -customer_name, -city) %>%
        group_by(project_name, product_name) %>%
        mutate(item_number = 1:n())
      ) %>%
    left_join(service_type) %>%
    mutate(product_type = case_when(
      product_name == '保温滑升门' ~ "insulated door",
      product_name == "液压踏知平台" ~ "dock lever",
      product_name == "铝制单板滑升门" ~ "aluminum door"
      )) %>%
    mutate(status = ifelse(!is.na(service_date), 1, 0),
           days = ifelse(
             !is.na(service_date),
             as.numeric(as.Date(service_date) - as.Date(install_date), units = "days"),
             as.numeric(as.Date("2021-04-22") - as.Date(install_date), units = "days")
             )) %>%
    mutate(man_made_damage = ifelse(is.na(man_made_damage), 0, man_made_damage))  %>%
    mutate(out_of_warranty = case_when(
      service_date > warranty_end_date & man_made_damage == 0 ~ 1,
      service_date <= warranty_end_date & man_made_damage == 0 ~ 0,
      is.na(service_date) & man_made_damage == 0 ~ 0
      ))

  cat("Adding customer groups.", fill = T)
  customer_groups <- create_customer_groups(service_data, cutoff = 100)
  customer_groups <- customer_groups$customer_groups %>%
    select(product_name, customer_type, customer_group)

  service_data <- service_data %>%
    inner_join(customer_groups) %>%
    distinct

  cat("Returning data.", fill = T)
  return(service_data)
}
```

Figure 7. Code of function

First, the implementation with data acquisition is initialized. To do this in R, survival curve, ggplot, tidyverse packages are needed [13]. The data type is in the csv format where 21 columns are presented and in total there are 5296 rows of data. To format this data R based fundamental coding and tidyverse is used, where the data is formatted by grouping by three basic products and formatting the date types according to the data given. Through calculation, having a dataset for the number of dates the product lasted for the survival curve later on can be achieved. Also, any man-made damage or existing warranty services was taken out since they do not have a proper cost for later calculations.

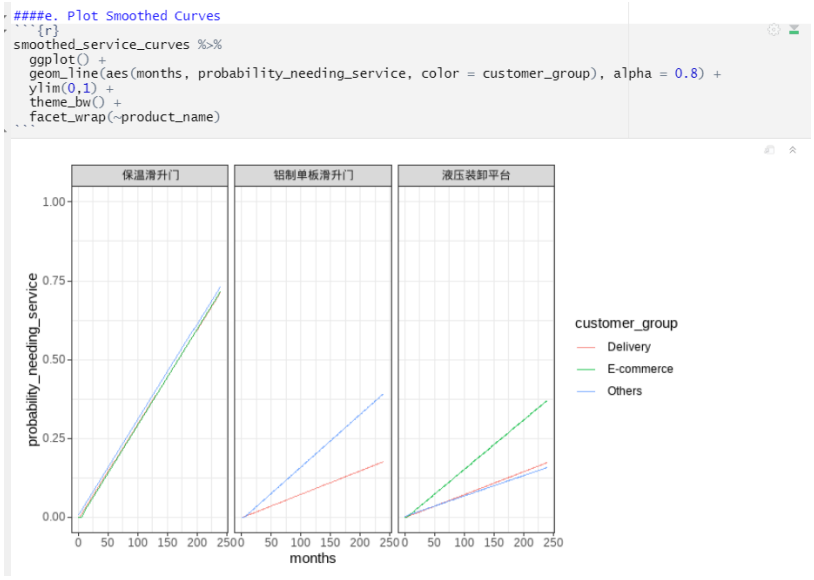


Figure 8. Code of plot smoothed curves

After getting and cleaning the data, the next step is to do some plotting the data to find the trend. a survival curve is used to generate a plot (figure 8) of probability of breakdowns during months according to different types of products and customers. Figure 8 is the result after ggplotting it and smoothing the curve. It will be later prepared with the average cost of services to calculate the recommended price of warranties.

```r
# forecast service costs
forecast_service_needs <- function(smoothed_curves,
                                   number_products = 100,
                                   financial_data){
  smoothed_curves %>%
    mutate_at(vars(product_name, customer_group), trimws) %>%
    mutate(total_products_needing_service = probability_needing_service * number_products) %>%
    mutate(new_products_needing_service = lead(total_products_needing_service) -
total_products_needing_service) %>%
    left_join(financial_data %>%
                mutate_at(vars(product_name, customer_group), trimws) %>%
                select(customer_group, product_name, avg_cost, avg_income)) %>%
    mutate(total_cost = total_products_needing_service * avg_cost,
           total_income = total_products_needing_service * avg_income) %>%
    select(-avg_cost, -avg_income) %>%
    mutate(total_profit = total_income - total_cost)
```

Figure 9. Code of forecast service costs

Then, this is a function that produces a data table that has values of predicting services needed for a hundred products, the percentage the average cost, income, and profit of the product combining with the outcome of the smoothen survival curve just produced.

```r
# calc recommended warranty

calc_recommended_warranty <- function(service_data = service_needs,
                                      financial_data = avg_financial_data,
                                      n_products = 100,
                                      profit_percentage = 1){

  service_data %>%
    mutate(year = ceiling(months / 12)) %>%
    group_by(product_name, customer_group, year) %>%
    summarize(total_probability_needing_service = max(probability_needing_service)) %>%
    mutate(n_total_products = n_products,
           projected_total_appointments = n_products * total_probability_needing_service) %>%
    mutate(appointments_per_year = lead(projected_total_appointments) - projected_total_appointments) %>%
    na.omit %>%
    inner_join(financial_data %>% select(customer_group, product_name, avg_cost)) %>%
    mutate(total_avg_running_cost = projected_total_appointments * avg_cost,
           avg_yearly_cost = appointments_per_year * avg_cost) %>%
    ungroup %>%
    select(product_name, customer_group, n_total_products, appointments_per_year, avg_yearly_cost) %>%
    distinct(product_name, customer_group, .keep_all = T) %>%
    rename(break_even_warranty_price = avg_yearly_cost) %>%
    mutate(recommended_warranty_price = break_even_warranty_price + (break_even_warranty_price *
profit_percentage),
           warranty_profit = recommended_warranty_price - break_even_warranty_price)


}
```

Figure 10. Code of recommended warranty

Eventually, a recommended warranty based on the profit percentage as an input is calculated. Based on the previously calculated cost, this code to produce a table of probability of breakdown and the trend of cost, income, profit by the different products and different customer groups is used.

4. EXPERIMENT

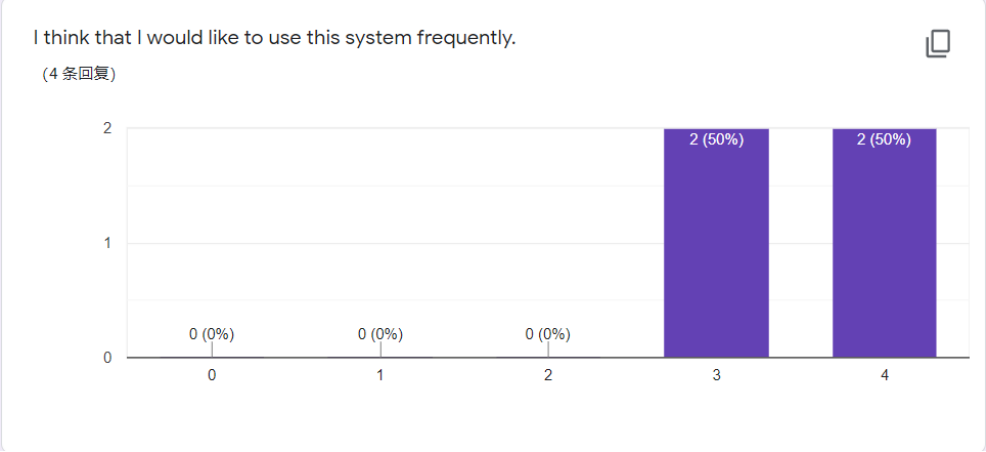


Figure 11. Survey result 1

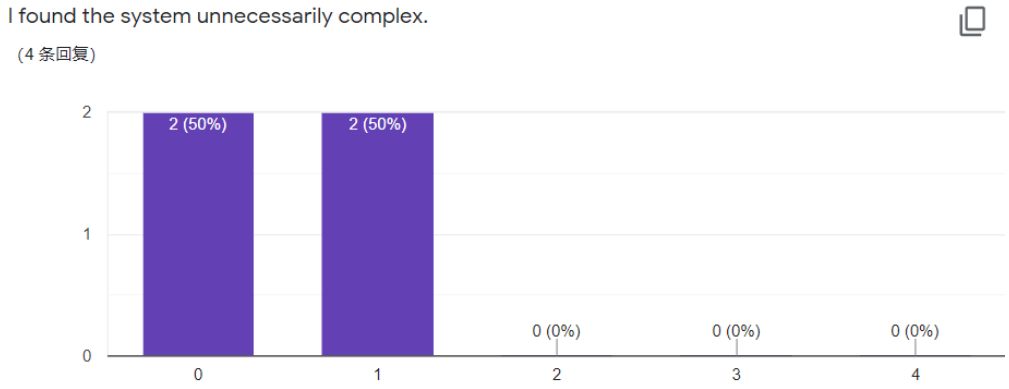


Figure 12. Survey result 2

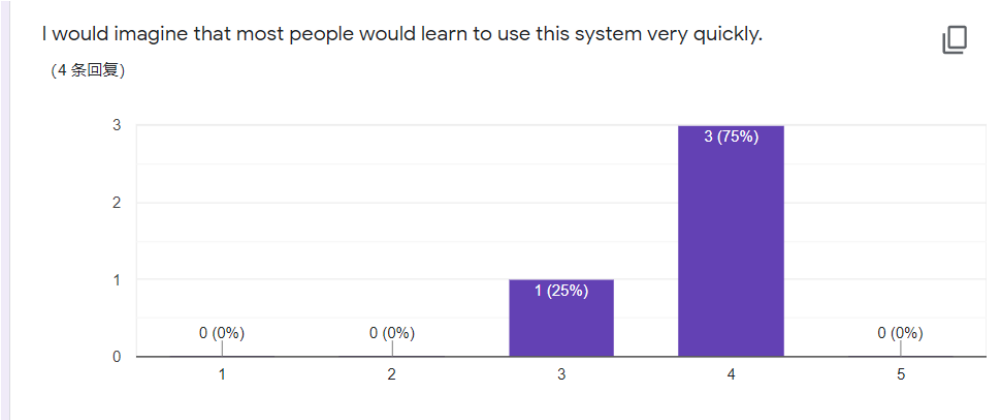


Figure 13. Survey result 3

The percentage of breakdown for each product is calculated. Including product types and business types as categorical variables, group by is used to filter out the history data that is specific to the product type and business types. After, the dates between the date the product are calculated that needs service to the date the product starts functions. Applying these variables into the survival model, which is a model that is usually used to calculate warranty for insurance companies, the regression line can be find that predicts the increase of probability in months. One interesting observation from the result is illustrated in figure 12, where the users rated the system based on its complexity. Since the overall system only has few pages, this question received a low score. This is output as how many services an order may need in x number of months in the table by multiplying the probability and the number of products in this order.

From this output, a test dataset can be use to find out if this data is correctly predicting in the real world situation.

5. RELATED WORK

A warranty forecasting model based on piecewise statistical distributions and stochastic simulation under the circumstance of having a large amount of data for the services already. Giving a interesting methodology to a similar question [14].

Another interesting domain in the field is the xxx presented by xxx in 2000. In general this paper forecasts the number of warranties through two phases. In phase I, they find upper and lower bounds of the warranty claim rates [15]. In phase II, they forecast for the recently launched product through the bounds in Phase I with a model built with the NHPP (non-homogeneous Poisson process) and the constrained maximum likelihood estimation.

This paper presents a forecasting method to predict a service system's expected number of through observed data which is used to calibrate a Generalized Renewal Processes (GRP) model [16]. It goes into detail of how the production in different months may impact the possibility of failures in the cars.

6. CONCLUSIONS

Warranty pricing is important to many businesses in different industry. For example, cars and headphones. Since the demand of having a warranty is huge, addressing the problem of how to price the warranty arises. Initially, initialized from a problem in a conversation, the research propose to bring a solution to the concern. Looking at the data from the department, an appropriate approach is identified after a conversation with the collaborator. A variety of solutions or proposals to solve this problem come up during the process. However, the survival curve is chosen because it fits the demand the best and it has been used in warranty companies. Later, the connection between the price and curve to produce the solution is last step needed.

Beginning with cleaning data, identifying different types of products, the subtraction between dates, and what data is man-made damage is needed since these data should not count. Before forecasting a recommended warranty price, the cost of services of products during a period is predicted. Using survival curves and ggplots, the prediction of the probability of breakdown is found. Combining the curve and costs, a recommended price of warranty to the users can be given eventually.

After all, user surveys as the experiment are used since not only the data has a strong correlation already in the plot, the existing warranty price does not give a good measurement on if the price is correct since it is pure guessing because the existence of the project is to help the company determine a correct price. Indeed, the experiment shows that the solution is effective and will help the company to envision its future success. Then, this is a successful data science project.

Current limitation is that the data sample is not enough. Since the service history system just came out two years ago for the company. Two year’s data cannot be adequate enough to predict all the trends after dividing them into different groups by business types and product types. Then the practicability of this model is being doubted.

Admittedly the data sample is too small, it can be solved very soon since FastLink is a fast growing company that has an enormous amount of data coming daily. Indeed, this new data can be used to optimize this model to predict better.

The system’s usability can be tested in the future. Through comparison between the date products actually breaks down in the future to see if the model works. If there is new data, the train dataset can be adjusted to see which will have the least error with the new incoming data. It will be a repeating process of iteration and optimization.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Jonathan Fei and Lina Tang for this opportunity to work on their data. Also, I'm deeply indebted to I would also like to extend my deepest gratitude to Beau Walker who helps me during the process of writing the code

REFERENCES

[1] Saltz, Jeffrey, and Kevin Crowston. "Comparing data science project management methodologies via a controlled experiment." (2017).
[2] Blischke, Wallace, ed. Warranty cost analysis. CRC Press, 2019.
[3] Newman, William M., and Michael G. Lamming. Interactive system design. Reading: Addison-Wesley, 1995.
[4] Hahn, Norbert, and R. Holzhauer. "Comparing dock levelers." Plant Engineering 50.11 (1996): 64-67.
[5] Johnson-Laird, Philip N., and Joanna Tagart. "How implication is understood." The American Journal of Psychology 82.3 (1969): 367-373.
[6] Freese, Frank. "Testing accuracy." Forest Science 6.2 (1960): 139-45.
[7] Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.
[8] Cameron, A. Colin, and Frank AG Windmeijer. "An R-squared measure of goodness of fit for some common nonlinear regression models." Journal of econometrics 77.2 (1997): 329-342.
[9] Wickham, Hadley, and Maintainer Hadley Wickham. "The ggplot package." Google Scholar. http://ftp. uni-bayreuth. de/math/statlib/R/CRAN/doc/packages/ggplot. pdf (2007).
[10] West, Kenneth D. "Forecast evaluation." Handbook of economic forecasting 1 (2006): 99-134.
[11] Baudrillard, Jean. The vital illusion. Columbia University Press, 2000.
[12] Lingis, Alphonso. The imperative. Indiana University Press, 1998.
[13] Wickham, Hadley. "The tidyverse." R package ver 1.1 (2017): 836.
[14] Kleyner, Andre, and Peter Sandborn. "A warranty forecasting model based on piecewise statistical distributions and stochastic simulation." Reliability Engineering & System Safety 88.3 (2005): 207-214.
[15] Wu, Shaomin, and Artur Akbarov. "Forecasting warranty claims for recently launched products." Reliability Engineering & System Safety 106 (2012): 160-164.
[16] Koutsellis, Themistoklis, et al. "Warranty forecasting of repairable systems for different production patterns." SAE International Journal of Materials and Manufacturing 10.3 (2017): 264-273.

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

AN IMPROVED FRAMEWORK FOR C-V2X SYSTEMS WITH DATA INTEGRATION AND IDENTITY-BASED AUTHENTICATION

Rui Huang

Lianyungang Jierui Electronics Co., Ltd.,
city: Lianyungang, zip code: 222061, China

ABSTRACT

Current trends of autonomous driving apply the hybrid use of on-vehicle and roadside smart devices to perform collaborative data sensing and computing, so as to achieve a comprehensive and stable decision making. The integrated system is usually named as C-V2X. However, several challenges have significantly hindered the development and adoption of such systems. For example, the difficulty of accessing multiple data protocols of multiple devices at the bottom layer, and the centralized deployment of computing arithmetic power. Therefore, this work proposes a novel framework for the design of C-V2X systems. First, a highly aggregated architecture is designed with fully integration with multiple traffic data resources. Then a multi-level information fusion model is designed based on multi-sensors in vehicle-road coordination. The model can fit different detection environments, detection mechanisms, and time frames. Finally, a lightweight and efficient identity-based authentication method is given. The method can realize bidirectional authentication between end devices and edge gateways.

KEYWORDS

Network Protocols, Wireless Network, Mobile Network, Virus, Worms & Trojon.

1. INTRODUCTION

Today's technology has made important progress in many fields and shows a cross-fertilization trend. In the integration of transportation systems, with the development of the Internet, a new generation of information technology represented by cloud computing, Internet of Things technology, intelligent sensing / big data mining technology is effectively integrated and applied to rail transportation, road transportation, water transportation and air transportation systems. This makes the integration of transportation systems show the trend of intelligence, networking and collaboration. At present, most of the world's major autonomous driving technology routes are solutions with the car as the intelligent body, i.e., the car itself is made into a mobile intelligent body. Such a solution has high technical requirements, and the system equipment is extremely expensive. This leads to its safety, reliability improvement of the input and output is relatively low as well as autonomous driving is difficult to be widely promoted in a short period of time, thus making it difficult to obtain the benefits of traffic efficiency and traffic safety. In addition, intelligent transportation application scenarios are complex and diverse. In the actual application process, a single public cloud or private cloud solutions are often difficult to meet the needs of intelligent transportation development.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 57-73, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111905

Against the background of the difficulty of enhancing single-vehicle autonomous driving technology and the increasing complexity of the traffic environment, autonomous driving increasingly relies on the development of intelligent road facilities [1]. C-V2X [2] vehicle-road cooperative system can realize different degrees of information interaction and sharing between vehicles and vehicles, vehicles and people, and vehicles and road traffic facilities by building roadside systems with sensing, fusion, path planning, control and communication functions, and vehicles only need to deploy low-cost on-board equipment. It is possible to have autonomous driving capability. This can lower the threshold of self-driving vehicles and shorten the time to realize large-scale autonomous driving, shortening the event of large-scale autonomous driving popularity by 10 to 15 years. Vehicle-road cooperative autonomous driving systems also consider different levels of cooperative optimization of vehicle-road distribution to efficiently and cooperatively perform vehicle and road sensing, prediction, decision making, and control functions.

Vehicle-road cooperative autonomous driving is a low-to-high development process, which mainly includes the following development stages

(1) Information interaction and collaboration, realizing information interaction and sharing between vehicles and roads. Using advanced wireless communication and new generation Internet and other technologies to realize dynamic real-time information interaction and sharing between vehicles and vehicles, vehicles and roads in all aspects, which is mainly reflected in the level of collection and fusion of environmental information by system participants.
(2) Perception prediction decision collaboration, on the basis of (1), to achieve vehicle-road collaboration perception and prediction decision function. With the saturation of vehicle technology progress space and the increase of traffic environment complexity, in addition to real-time information interaction and sharing with the help of communication technology, the realization of autonomous driving perception and decision making also depends on intelligent road facilities and in-vehicle equipment such as radar and cameras. The above facilities and equipment are used to realize the sensing of dynamic traffic environment information in all-time and space, as well as the subsequent functions of data fusion, state prediction and behavioral decision-making. This is mainly reflected in the comprehensive collection of environmental information by system participants and the driving decision level.
(3) Realize advanced vehicle-road cooperative control function. On the basis of (2), it can also realize the vehicle-road cooperative automatic driving control function, and then complete the full coverage of the whole key steps of automatic driving. For example, it can be applied in limited scenarios such as highway lanes, urban expressways and automatic parking, which mainly reflects the comprehensive collection of environmental information by system participants, driving decision and control execution at the whole level.
(4) The vehicle and the road achieve comprehensive synergy, i.e. complete system functions such as vehicle-road cooperative sensing, vehicle-road cooperative prediction and decision making, and vehicle-road cooperative control integration. It further enhances the intelligent role of road infrastructure, so as to realize the comprehensive intelligent collaboration and cooperation between vehicles and roads, i.e., to realize the system integration functions of vehicle-road cooperative sensing, vehicle-road cooperative prediction and decision making, and vehicle-road cooperative control in any scenario. Vehicle-road synergy improves the commercialization of vehicle autonomous driving and forms an integrated development path in which vehicles and roads jointly promote the realization of autonomous driving.

Baidu has brought together autonomous driving and road-vehicle collaboration. For vehicle intelligence, Baidu has launched and open-sourced the Apollo platform [3], which has attracted a large number of developers and manufacturers and has now been updated to Apollo 6.0. Baidu has launched the "ACE Traffic Engine" [4] to build a modern intelligent transportation system with real-time sensing, instantaneous response and intelligent decision-making. Currently, Baidu's "ACE Traffic Engine" integrated solution [5] has been put into practice in nearly 20 cities, including Beijing, Changsha and Baoding. Compared with Baidu, Alibaba is more concerned about the control platform of vehicle-road coordination. Its proposed ET City Brain [6], together with AliOS on the vehicle side, provides global analysis and scheduling at the city level, and has already achieved milestones. In Beijing, through signal timing optimization, the average delay of motor vehicles through intersections has dropped by 6% and the parking ratio has been reduced by 3%. In Shanghai, the prediction accuracy of the neural network model built for the traffic status of the north-south elevated sections has improved by 10% [7].

Vehicle-road cooperative intelligent transportation faces many technical challenges and development bottlenecks, such as the difficulty of accessing multiple data protocols of multiple devices at the bottom layer, and the centralized deployment of computing arithmetic power cannot meet the demand for computing latency of intelligent applications. In the intelligent vehicle-road cooperative system, there are many kinds of sensors and wide distribution, and the in-vehicle sensing system is still in high-speed motion, and the detection environment, detection mechanism, time base, information characteristics and description methods of sensors are different. The problem of fusion of sensor information [8] from multiple locations prevails. In the face of ultra-intensive data volume, as well as the status quo of low tolerance to time delays, existing methods are difficult to play the advantages and characteristics of both sides of the cloud, in order to ensure the safety of road traffic system, usually with the premise of efficiency, to improve the efficiency of road control. We propose a C-V2X vehicle-road collaboration system for road traffic environment, build a vehicle-road collaboration system architecture based on V2X vehicle-road collaboration wireless communication, multi-access edge computing and high-precision positioning, and form a comprehensive solution for city-level vehicle-road collaboration intelligent transportation.

Our contributions:

1. We propose a new vehicle-road collaboration architecture, which fully integrates multiple traffic data resources, as well as control resources such as traffic monitoring, guidance screens, and signal control from three levels: individual vehicles, intersection localization, and regional road network.
2. We propose a multi-level information fusion technology based on multi-sensors in vehicle-road coordination to realize multi-sensor information fusion based on vehicle-road coordination perception, in view of the characteristics of many types of sensors and wide distribution in the vehicle-road coordination system, and the different detection environments, detection mechanisms, time frames, information characteristics and description methods of sensors.

We propose a lightweight identity-based authentication method to improve the authentication protocol for identity-based authentication and realize bidirectional authentication between end devices and edge gateways.

2. RELATED WORK

Existing autonomous driving technologies can be divided into two major technical routes: single-vehicle intelligence and vehicle-road collaboration. Single-vehicle intelligence relies entirely on the input of information from on-board sensors (such as LIDAR, millimeter wave radar and cameras) for environmental perception, and then artificial intelligence technology for environmental change prediction and driving decision generation, such as Waymo, Tesla, Mobileye and other companies in the United States have achieved L2-L4 level autonomous driving to some extent. However, in bad weather such as night, fog, rain and snow or complex traffic scenarios such as intersections and curves, the accuracy and reliability of on-board sensors are difficult to guarantee, thus making it impossible to achieve autonomous driving.

2.1. Single Vehicle Intelligence

The Society of Automotive Engineers (SAE) has defined five levels of driving automation. In this taxonomy, level zero represents no automation at all. Primitive driver assistance systems, such as adaptive cruise control, antilock braking systems and stability control, start at Level 1. Level 2 is partial automation, into which advanced assistance systems such as emergency braking or collision avoidance are integrated. The third level is conditional automation. During normal operation, the driver can focus on tasks other than driving, however, he/she must respond quickly to emergency alerts from the vehicle and be ready to take over. No level of human attention is required at Levels 4 and 5. However, Level 4 can only operate in limited ODDs where special infrastructure or detailed maps exist. In the case of leaving these areas, the vehicle must stop the trip by automatically stopping. A fully automated system, Level V, can operate on any road network and in any weather conditions. There are no production vehicles capable of Level 4 or Level 5 automation.

Self-driving cars use sensors such as cameras, radar and LIDAR to sense their surroundings. Due to the high price of LIDAR, Tesla wants to achieve fully autonomous driving without the use of LIDAR and proposes Autopilot [9] to be loaded on all Tesla production cars in the future. autopilot has achieved L3 level of autonomous driving by collecting image data through 8 cameras and 12 millimeter wave radars to assist in perception, but it is difficult to continue to improve. Waymo, which ranks first in the world in the field of autonomous driving, is mainly dedicated to the research of autonomous driving algorithms, building its own radium map through short-range, medium-range and long-range lidar, and choosing electromagnetic wave radar with better environmental adaptability than ultrasonic radar, but the requirements for cameras are also higher.

Baidu's Apollo has now achieved L4 level autonomous driving capability in a semi-enclosed environment. In the latest versions of Apollo, it is gradually moving closer to V2X, making Apollo with high level autonomous driving capabilities even more advantageous. Tencent, on the other hand, is more focused on providing services [10] for autonomous driving with the help of high-precision maps [11].

2.2. Cellular-Vehicle to everything

The Internet of Vehicles includes Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I), Vehicle to Pedestrian (V2P), and Vehicle to Network (V2N) interactions. Network (V2N) interaction. Telematics will rely on information and communication technology to provide comprehensive information services through all-round connection and data interaction, forming a

new industrial form with deep integration of automobile, electronics, information and communication, road transport and other industries.

The V2X network architecture based on vehicle-road cooperation can be generally divided into three levels: perception layer, decision layer and execution layer. The perception layer mainly involves environment perception technology and vehicle positioning technology to obtain the location of the vehicle and the surrounding traffic status; the decision layer mainly involves environment change prediction technology and driving decision technology, i.e., to predict the movement trajectory of the surrounding people and vehicles and generate optimized driving decisions accordingly; the execution layer mainly performs driving decisions through mechanical control. In the V2X network architecture, cloud, edge-side and vehicle-side are included, and each part is involved in various aspects of smart transportation, including data sensing, analysis and simulation, and decision control. The cloud has the lowest real-time performance, the edge side is in the middle, and the vehicle side has the highest real-time performance. Quasi-real-time data fusion and downlink in the edge cloud enables driverless vehicles to gain the ability to acquire information that breaks through visual dead spots or across occlusions. The perception, policy and control related to real-time vehicle control are performed at the vehicle side, and the analysis and calculation are done by the cloud-side fusion control platform.

2.3. Communication technology for C-V2X

C-V2X defines V2X technology based on cellular communications, including LTE-V2X, 5G-V2X [12]. it leverages the already existing LTE network facilities to enable V2V, V2N, V2I information interaction. The most attractive aspect of this technology is its ability to keep up with changes, adapt to more complex security application scenarios, meet low latency, high reliability and satisfy bandwidth requirements. LTE V2X defines two communication methods for vehicle applications [13]: centralized (LTE-V-Cell) and distributed (LTE-V-Direct). The centralized type, also known as cellular type, requires a base station as the control center, which defines the communication between the vehicle and roadside communication unit and base station equipment; the distributed type, also known as direct type, does not require a base station as support. 2006's CoCar project achieved an end-to-end delay of less than 500ms [14]. 2017's LTE-V2X technology from Bosch and Huawei achieved direct communication coverage of 1km and more. above, which can effectively provide the performance of two cars following each other face-to-face at 500km/h, with communication latency less than 20ms in high-density congested traffic scenarios and message sending success rate over 90%.

5G-V2X is the V2X standard for 5G communication. Because 4G-LTE technology was not fully considered at the beginning of the design, and with the rapid development of smart cars, 4G-LTE technology [15] became insufficient. V2X will be part of the 5G network, and 5G-V2X has the potential to integrate LTE-V2X and DSRC [16] to provide safer and more efficient operation capabilities for cars [17]. In July 2020, 3GPP completed the first 5G framework-based 5G-V2X standard Rel-16, which realizes the cooperative sensing and path planning, thus effectively avoiding accidents.

2.4. C-V2X security issues

Automated vehicles in vehicle-road collaboration interact with the surrounding environment through wireless communication technology [18], so it is of great significance to achieve a secure and reliable communication method to guarantee the security of Telematics [19][20]. The literature [21] proposes a security scheme for Telematics communication based on public key architecture and edge computing. The authors use the location information of the vehicle as well as the RSU as the reference basis for key pair distribution, so that key pre-distribution can be

performed based on the prediction of the vehicle location. However, the security assurance scheme based on public key architecture requires certificate delivery and verification during the communication process, which greatly increases the load on the network. In the literature [22], an efficient anonymous, bulk authentication scheme is proposed to address the problem of security privacy protection in vehicular networking, thus reducing the message loss rate of vehicles and RSUs.

3. SYSTEM MODEL

We propose a C-V2X vehicle-road collaboration system for road traffic environment, which is divided into a central cloud platform, an edge computing system, roadside devices and terminal devices.

The C-V2X vehicle-road collaboration system for road traffic environment defines the architecture and functional requirements of the central cloud platform and the edge computing system, and proposes the improvement of multi-sensor fusion and safety warning technology based on high-precision positioning.

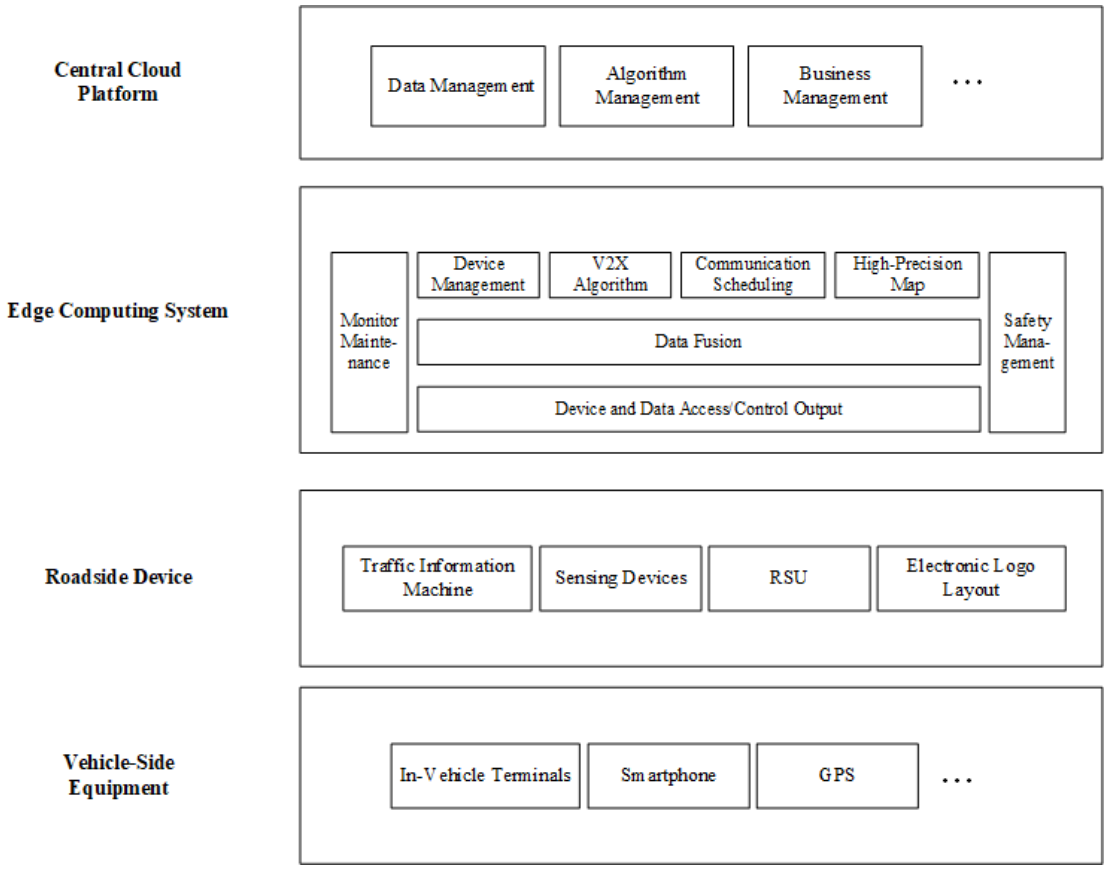


Figure 1. System Architecture

3.1. Central cloud platform

The core of C-V2X vehicle-road coordination system in the central platform assumes the functions of macroscopic decision-making and unified command dispatch. By converging the panoramic sensing data of people, vehicles, roads and environment in the vehicle-road

cooperation scenario, it provides support for vehicle-road and vehicle-brain, and vehicle-vehicle cooperation decision based on big data and AI.

Distributed object storage enables storage of all data types, solving the problems associated with storage of massive amounts of all data forms, and making it more available, fault-tolerant, and scalable than a single data center. Make the system scalable by referring to objects using IDs instead of filenames. Associate large amounts of metadata with specific objects. Perceptual device layer convergence access to the cloud computing platform to provide basic data support for the cloud computing platform.

The functions of the central cloud platform include:

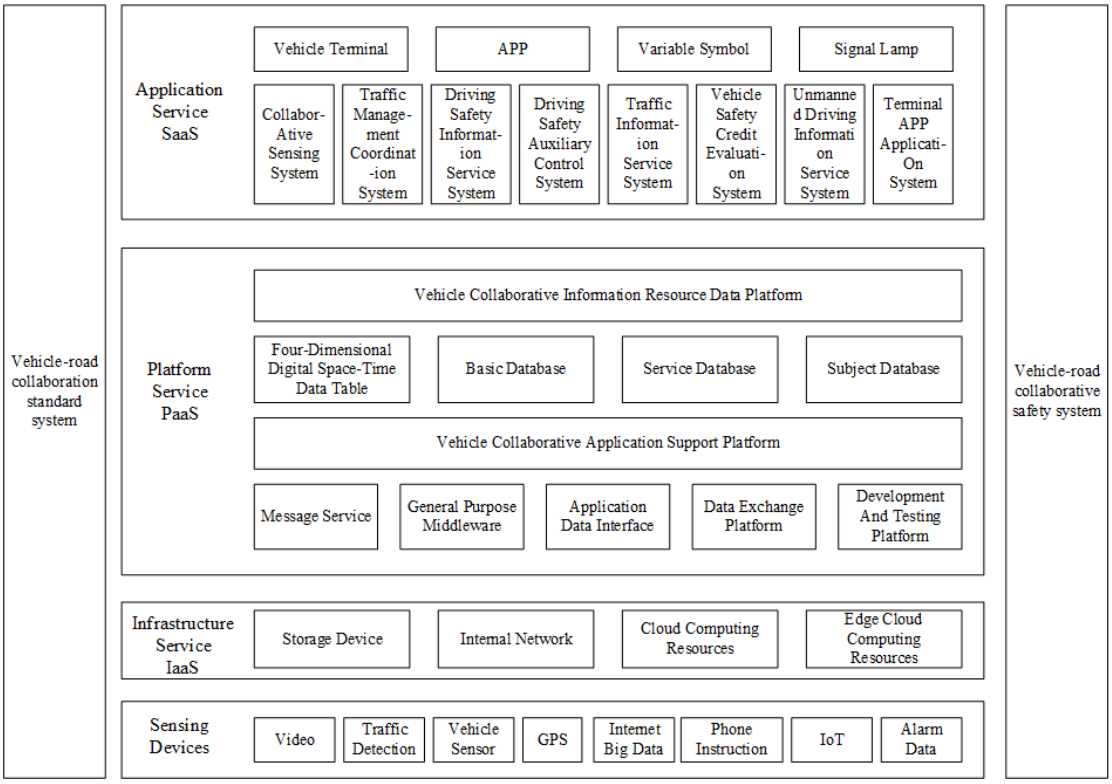


Figure 2. Central System Architecture

(1) Data management: Aggregate a huge amount of front-end devices, including pictures, videos, status and other timing data of vehicles, traffic signal control machines and many other devices. Distributed storage architecture is adopted for persistent storage of massive data.

(2) Data fusion: Relying on the processing power of the big data platform and the integrated data processing algorithms and models, the multi-source traffic data are automatically analyzed and processed for optimization under the set rules. It provides data-level fusion, feature-level fusion and decision-level fusion services to complete the required traffic control decision and prediction and early warning.

(3) Edge cloud collaboration:

1) Security collaboration: Provide perfect security policies, including traffic cleaning, traffic analysis, etc. In the process of security policy collaboration, the center can block malicious traffic if it is found to exist at an edge to prevent malicious traffic from spreading throughout the edge cloud platform.

2) Application collaboration: The cloud realizes lifecycle management of value-added network applications at edge nodes, including application push, installation, uninstallation, update, monitoring and logging. The central node can incubate and start the already existing application images on different edge clouds to complete the high availability guarantee and hot migration of applications.

(4) GBA security authentication function: In addition to grouping terminals with high similarity, security authentication and device management of the cloud platform are decoupled and decentralized for deployment to the edge gateway authorized through authentication. The designed lightweight certificate-free authentication protocol proposes a distributed authentication mechanism with the edge gateway as the core, thus improving the authentication efficiency.

Device management: centralized management of various devices accessed by the system, including vehicles, guidance screens, traffic signal controllers, and sensors.

3.2. Edge System

Edge computing system is the roadside core system of C-V2X vehicle-road cooperation system for road traffic environment, and it is the main undertaking system for communication and authentication services of C-V2X.

The edge computing system is deployed at the edge side of the front end of the vehicle-road collaboration system near the traffic road and traffic data sources. This system incorporates an open platform of network, computing, storage, and application core capabilities, and provides computing and intelligence services. It extend cloud computing and intelligence capabilities to edge nodes close to end devices. This avoids the problems of longer network latency, network congestion, and degraded quality of service that can be caused by putting computing on the cloud. This satisfies the requirements of high real-time and high computing capability of the vehicle-road collaboration system. The structure is shown in Figure 3.

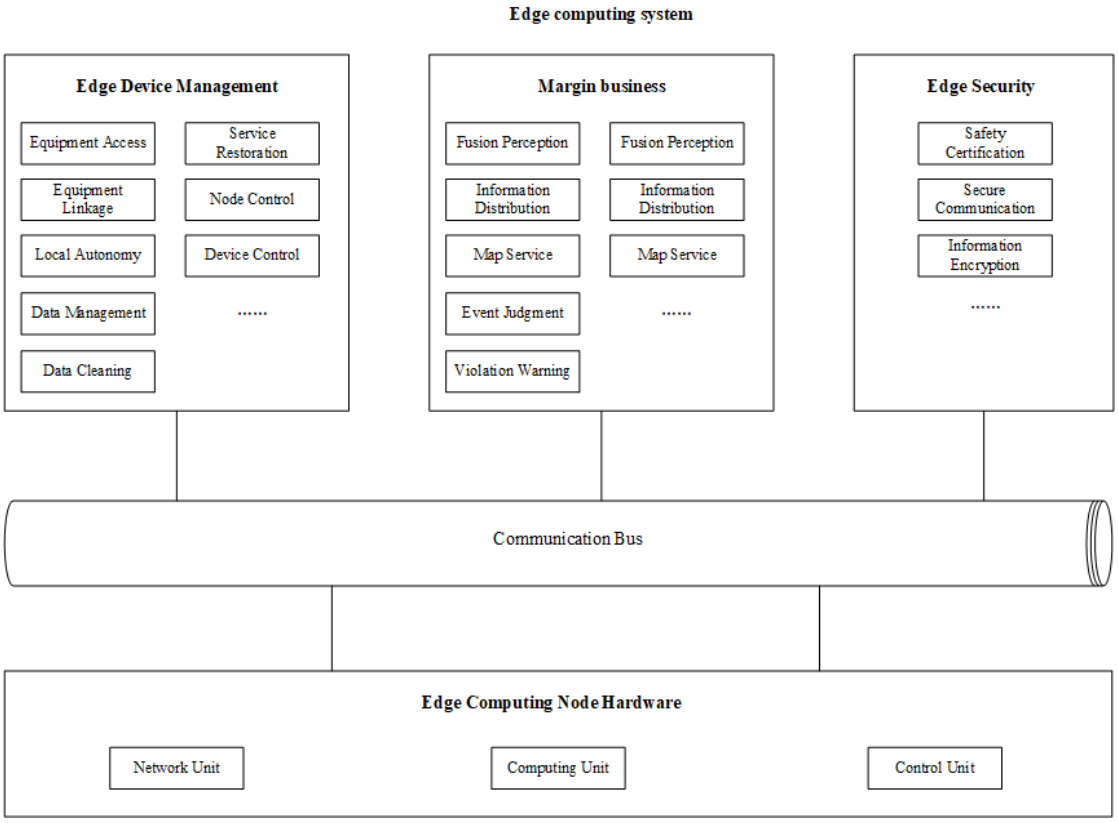


Figure 3. Edge Computing System Architecture

The edge layer includes two main parts, edge nodes and edge management. The edge node is the hardware entity, which is the core of carrying edge computing services. The presenting core of edge management is software, and the main function is to provide unified management of edge nodes.

The physical composition of the edge computing nodes for vehicle-road collaboration includes three basic modules: network, computing and storage. The service execution of C-V2X vehicle-road cooperative edge computing for road traffic environment is inseparable from the support of communication network. The network of edge computing is characterized by the need to satisfy both the determinism of transmission time and data integrity of control-related services, and the ability to support flexible deployment and implementation of services. Time-sensitive network (TSN) and software-defined network (SDN) technologies will be important fundamental resources for the network part of edge computing. Heterogeneous computing is the key computing hardware architecture at the edge. Edge devices have to handle both structured data and unstructured data at the same time. A local database is used to store high-precision map data, high-resolution image data, etc. It supports functions such as fast writing of time-series data, persistence, and multi-dimensional aggregated queries.

The edge computing node of vehicle-road collaboration logically contains three functional units: control, analysis and optimization. The control function unit perceives the environment timely and accurately, and uses edge computing to enhance local computing capabilities and reduce the response latency caused by cloud-centric computing. The control functional unit mainly includes the functions of environment sensing and execution, real-time communication, entity abstraction, control system modeling, device resource management, and program operation executor. The analysis functional unit mainly includes streaming data analysis, video image analysis, intelligent

computing, and data mining. The algorithms such as neural network and machine learning related to artificial intelligence are applied at the edge side to complete the solution of complex problems using intelligent computing.

The integration of edge computing system with C-V2X can enhance the end-to-end communication capability of C-V2X. It can also provide auxiliary computing and data storage support for C-V2X vehicle-road cooperation application scenarios for road traffic environment.

The GBA secure communication system runs in a multi-access edge computing system. It provides complete security authentication and session channel encryption services for application layer services such as vehicle networking, road infrastructure networking, and vehicle-road collaboration. In the C-V2X vehicle-road cooperation system for road traffic environment, GBA secure communication connects the vehicle terminal and USIM card with the wireless access network, core network, GBA platform, and CA server. This guarantees the communication security of V2X.

3.3. Traffic information sensing module

Traffic targets in the vehicle-road cooperative environment refer to multi-source multi-dimensional traffic information such as vehicles, pedestrians, signage, and intersection environment. Multi-source multidimensional traffic information acquisition is achieved by sensors such as radar detectors, video detectors, and small meteorological data collectors. We propose a multi-channel intermingled information fusion method based on radar and video to achieve more accurate traffic target recognition by fusing traffic information from multiple sensors.

The point of video vehicle detection is that it can provide the impact situation of the road surface, and has a more excellent detection performance for beating cars and dense small cars in the image area under ideal lighting conditions. The recognition of different vehicle models and pedestrians is also good. But video vehicle detection is greatly affected by the light situation and weather conditions. For example, the recognition performance at night becomes poor, rain and fog weather basically cannot work, and adhere to the limited distance, generally not more than 100 meters.

Radar vehicle detection can detect a section of the road and can quickly derive precise position and speed information of the target, which is less affected by weather conditions. However, radar vehicle detection cannot give image information, cannot detect stopping targets, is not effective for hitting vehicles, and cannot effectively distinguish between various types of vehicles and pedestrians.

The fusion of traffic information from video vehicle detection and radar vehicle detection has two advantages, respectively: first, a sufficiently large detection area, sufficiently accurate data, and unaffected by climatic conditions is guaranteed by radar. Second, it can provide image information by video detection to accomplish stationary targets, large vehicle detection and distinguish various vehicle categories.

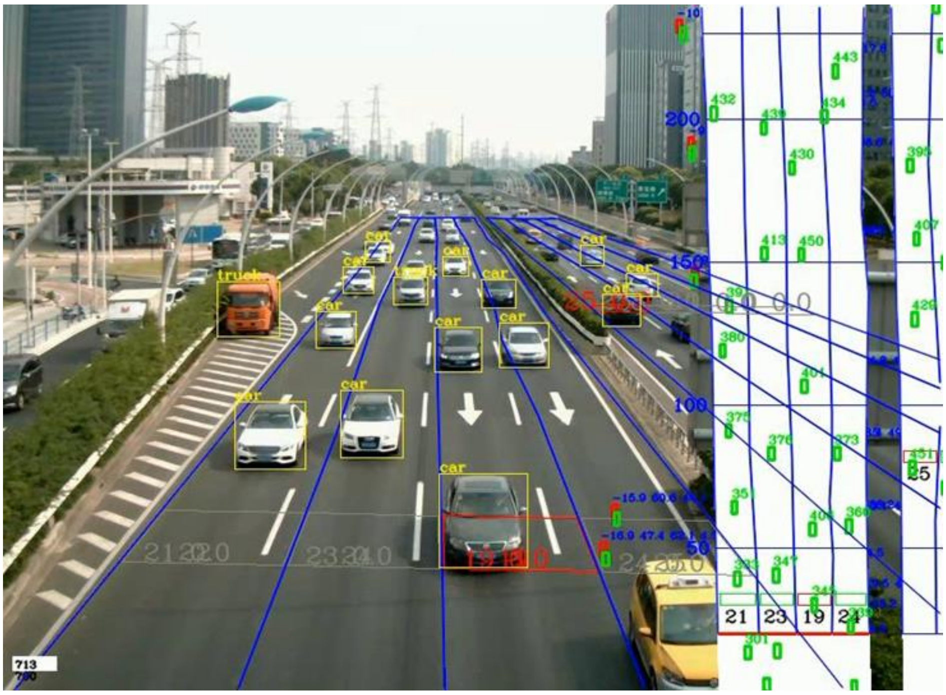


Figure 4. Radar and Video Traffic Information Fusion Perception

For the same target, both radar and video detect the target with high confidence. Both radar and video detect the target, but with high or low confidence. Only one of the radar and video detects the target. Only one of the radar and video detects this target. Neither radar nor video detects this target. The target data obtained by using both detection methods are weighted by confidence to obtain the final target data, and the final target data is used to update the target tracking status.

Since the type, style, and clarity of data collected by various other sensors are different, it is first necessary to perform spatial coordinate system conversion and temporal calibration on these multidimensional data. After the calibration, the fusion of radar video data with the same time stamp in the same coordinate system is achieved. The fusion centers after spatio-temporal alignment appear in clusters and are roughly distributed around the target true value. According to this characteristic using the idea of clustering in pattern recognition theory, the data belonging to the same target in different detectors can be clustered into one class and the data that are not targets can be separated. The video centers clustered into one class are estimated to calculate the fusion center, so that the tracking of multi-dimensional traffic targets is converted into a single video multi-target tracking problem.

The converted but video target data is input to the neural network. The MobileNet-SSD model is used as the core algorithm of the detection module. the basic structure of the SSD vehicle detection model is shown in Figure 5.

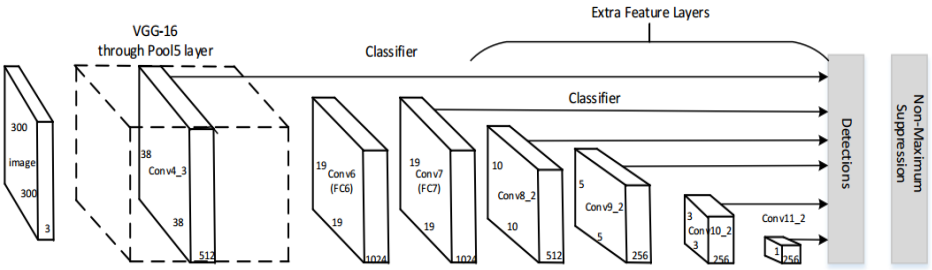


Figure 5. SSD vehicle inspection model

Firstly, the video to be detected by vehicle is pre-processed frame by frame so that the network input is an image of size 300*300*3. 300 denotes the number of pixels of image length and width, and 3 denotes the three channels R, G, and B. The SSD uses a feature pyramid structure for detection. Its base network structure is VGG-16, and the last two fully connected layers are changed to convolutional layers, and subsequently four convolutional layers are added to construct the network structure. The decreasing size of the convolutional layers layer by layer enables the network to make predictions at multiple scales. The convolutional layers extract different vehicle features and the size of the output feature map becomes smaller layer by layer. The size of the feature image obtained from each layer is 38*38, 19*19, 10*10, 5*5, 3*3, and 1*1, respectively.

For each additional feature layer added, a series of convolution kernels are used to perform convolution operations by means of sliding windows to produce the corresponding predictions. The output (feature map) of each of the five different convolutional layers is convolved with two different 3x3 convolutional kernels. One output is classified using confidence, with each default box generating 2 category confidences, and the other output is regressed using localization, with each default box generating 4 coordinate values (x,y,w,h). Specifically for a feature layer of size mxn with p channels, using a small convolution kernel of 3x3xp, a score or coordinate offset with respect to the default box is generated at each position of mxn for the attributed category. These 5 feature maps also go through the PriorBox layer to generate the priority box (the default box selected in practice). The number of default boxes in each layer of the above 5 feature maps is given, and the total number of obtained prior boxes in all feature maps is 8732.

The position of each box is fixed with respect to the feature lattice corresponding to it. In each feature lattice, the displacement from the default box needs to be predicted and the score (probability of belonging to a certain class) of the objects contained in each box. So, for each box in a set of k boxes at a location, c classes, the score of each class, and 4 offsets of that box compared to its default box can be calculated. Therefore, (c+4)*k results are generated for each location in the feature map. For a feature map of size m*n, then (c+4)*k*m*n outputs are generated.

MobileNet-SSD is evolved from SSD. It is a lightweight deep network model proposed for application to mobile, replacing the VGG-16 base feature extraction network in the SSD network with the MobileNet network. It can significantly improve the computational efficiency with guaranteed accuracy. The main use of Depthwise Separable Convolution decomposes the standard convolutional kernel to reduce the computational effort. Depthwise Separable Convolution is a standard convolution kernel divided into a depthwise convolution kernel and a pointwise convolution kernel of 1*1. The ratio of kernel computation is: (DK*DK*M* DF*DF+M*N* DF*DF)/(DK*DK*M*N*DF*DF)=1/N+1/(DK*DK), the computation is greatly reduced.

The detection results directly output by the detection network will contain a large number of duplicate frames, here we use the non-maximum suppression algorithm to filter out the duplicate detection frames and get the vehicle detection results. The main idea of the non-maximum suppression algorithm is as follows: for any detection target, the redundant detection frames should be eliminated, leaving only the frames with the highest confidence. The final remaining window within the sequence is then outputted as the final detection result. Thus, the recognition of multidimensional targets is achieved.

3.4. Security Module

The identity-based authentication protocol is improved to address the problem of limited communication and computational resources of IoT end devices. The resulting authentication scheme with lower computational complexity and higher efficiency is proposed. This achieves bi-directional authentication of end devices and edge gateways.

The identity-based bidirectional authentication protocol is divided into three main phases: as shown in Figure 6

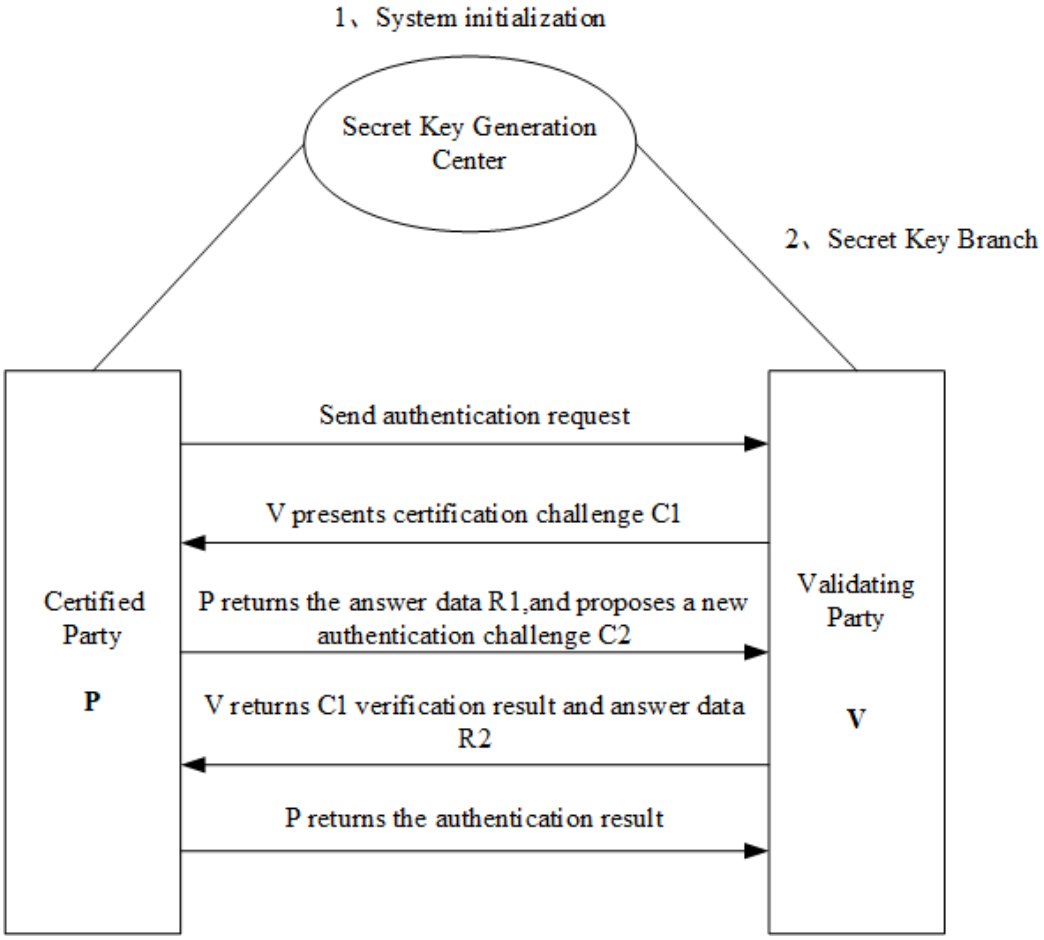


Figure 6. Two-way authentication protocol

(1) Initialization phase

At the completion of the private key generation center (PKG), the PKG picks the additive cyclic group G_1 and the multiplicative cyclic group G_2. Both G_1 and G_2 are prime q-order. P is an arbitrarily chosen generating element of G_1. Denote the bilinear pair as:

$$e: G_1 \times G_2 \to G_1$$

Given a security parameter 1^k, select the random number $sk_m \grave{o} Z_p^*$., sk_m is the master key (i.e., the system private key), and compute the master public key (i.e., the system public key).

$$P_{pub} = sk_m P$$

Select hash function.

$$h_1: \{0,1\}^* \to G_1$$

$$h_2: \{0,1\}^* \times G_2 \to Z_q^*$$

The master key cannot be made public by the PKG alone, and the system reference $\left(G_1, G_2, e, q, P, P_{pub}, h_1, h_2\right)$ is published to all users in the system.

(2) Key extraction stage

Completed in the PKG, the corresponding public key is calculated based on the unique identification ID of the user in the system:

$$Q_{ID} = h_1(ID)$$

The private key corresponding to the user is then generated from the master key:

$$sk_{ID} = sk_m Q_{ID}$$

After that, the user key pair (Q_{ID}, sk_{ID}) is sent to the corresponding user through a secure channel (online or offline).

(3) Two-way authentication phase

T is the authentication initiator (i.e., terminal device) and S is the authentication server (edge gateway). To improve security, two-way authentication of T and S is required.

(a) T selects the random number $r_T \grave{o} Z_q^*$, the timestamp t_T and the hash function h_2 to calculate

$$c_T = h_2(r_T, t_T)$$

(b) S decrypts c_T after receiving the data, first verifies the validity of the timestamp t_T, selects the random number $r_S \grave{o} Z_q^*$ when the detection result is a valid timestamp, and calculates

$$m_S = (r_S + r_T) sk_S$$
$$c_S = h_2(r_S, t_S)$$
$$Q_T = h_1(ID_T)$$

Re-transmitting authentication challenge information to the end device T.

(c) T decrypts c_S after receiving the data, verifies the validity of timestamp t_S, and if the timestamp is valid, calculates the public key of S.

$$Q_S = h_1(ID_S)$$

and test whether the equation holds:

$$e(m_S, -(r_S + r_T)P) = e(Q_S, P_{pub})$$

If the equation does not hold then it shows that T fails to authenticate to S, i.e., S is not a legitimate edge gateway for the end device and T disconnects from the connection; otherwise it proves that T authenticates to S successfully and calculates.

$$m_T = (r_S + r_T) sk_T$$

return an answer message to S upon completion.

d) S decrypts and verifies that the equation holds after receiving a successful message from T and verifying that the timestamp is successful

$$e(m_S, -(r_S + r_T)P) = e(Q_S, P_{pub})$$

If the equation does not hold, it indicates that S fails to authenticate T, the identity of the terminal device is considered illegitimate, a failure frame is returned, the data uploaded by terminal T is rejected and the connection is disconnected; otherwise, it indicates that S authenticates T successfully, a success frame is returned and the data uploaded by T starts to be received.

4. CONCLUSIONS

This paper proposes a novel framework for C-V2X system design. The framework can integrate multi-source data and perform collaborative analysis, which may benefit numerous services like autonomous driving. To achieve this goal, a highly aggregated architecture is proposed to hierarchically fuse the data resources. Then a multi-modal information fusion method is proposed to further aggregate the multi-sensor data generated within the C-V2X system. The method is flexible for different detection tasks and scenarios. Finally, the paper also introduces a fast and

reliable authentication method to enhance the security level of the whole system. In our future study, we will focus on the detailed methods and models used for our C-V2X systems.

REFERENCES

[1] Wang Y, Chao W L, Garg D, et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8445-8453.

[2] Chen S, Hu J, Shi Y, et al. A vision of C-V2X: technologies, field testing, and challenges with chinese development[J]. IEEE Internet of Things Journal, 2020, 7(5): 3872-3881.

[3] Fan H, Zhu F, Liu C, et al. Baidu apollo em motion planner[J]. arXiv preprint arXiv:1807.08048, 2018.

[4] Liu Z, Lu F, Wang P, et al. 3D Part Guided Image Editing for Fine-Grained Object Understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11336-11345.

[5] Du L, Ye X, Tan X, et al. Associate-3ddet: perceptual-to-conceptual association for 3d point cloud object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13329-13338.

[6] Yu Z, Liang S, Wei L, et al. MaCAR: Urban Traffic Light Control via Active Multi-agent Communication and Action Rectification[J].

[7] Chu W, Liu Y, Shen C, et al. Multi-task vehicle detection with region-of-interest voting[J]. IEEE Transactions on Image Processing, 2017, 27(1): 432-441.

[8] Zou L, Wang Z, Hu J, et al. Moving horizon estimation meets multi-sensor information fusion: Development, opportunities and challenges[J]. Information Fusion, 2020, 60: 1-10.

[9] Pan T, Song Y, Yang T, et al. Videomoco: Contrastive video representation learning with temporally adversarial examples[J]. arXiv preprint arXiv:2103.05905, 2021.

[10] Jia K, Kenney M, Mattila J, et al. The application of artificial intelligence at Chinese digital platform giants: Baidu, Alibaba and Tencent[J]. ETLA reports, 2018 (81).

[11] Ding L, Feng C. DeepMapping: Unsupervised map estimation from multiple point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8650-8659.

[12] Molina-Masegosa R, Gozalvez J. LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications[J]. IEEE Vehicular Technology Magazine, 2017, 12(4): 30-39.

[13] Chen S, Hu J, Shi Y, et al. LTE-V: A TD-LTE-based V2X solution for future vehicular network[J]. IEEE Internet of Things journal, 2016, 3(6): 997-1005.

[14] Gonzalez-Martín M, Sepulcre M, Molina-Masegosa R, et al. Analytical models of the performance of C-V2X mode 4 vehicular communications[J]. IEEE Transactions on Vehicular Technology, 2018, 68(2): 1155-1166.

[15] Vukadinovic V, Bakowski K, Marsch P, et al. 3GPP C-V2X and IEEE 802.11 p for Vehicle-to-Vehicle communications in highway platooning scenarios[J]. Ad Hoc Networks, 2018, 74: 17-29.

[16] Ghafoor K Z, Guizani M, Kong L, et al. Enabling efficient coexistence of DSRC and C-V2X in vehicular networks[J]. IEEE Wireless Communications, 2019, 27(2): 134-140.

[17] Boban M, Kousaridas A, Manolakis K, et al. Use cases, requirements, and design considerations for 5G V2X[J]. arXiv preprint arXiv:1712.01754, 2017.

[18] Wang J, Wu J, Li Y. The driving safety field based on driver–vehicle–road interactions[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2203-2214.

[19] Nassi B, Mirsky Y, Nassi D, et al. Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020: 293-308.

[20] Jarašūniene A, Jakubauskas G. Improvement of road safety using passive and active intelligent vehicle safety systems[J]. Transport, 2007, 22(4): 284-289.

[21] Zhang J M, Zhao Y J, Jiang H B, et al. Research on protection technology for location privacy in VANET[J]. Journal on Communications, 2012, 33(8): 180.

[22] Jing T, Pei Y, Zhang B, et al. An efficient anonymous batch authentication scheme based on priority and cooperation for VANETs[J]. EURASIP Journal on Wireless Communications and Networking, 2018, 2018(1): 1-13.

AUTHORS

Rui Huang, born in June 1984 in Xuzhou, Jiangsu Province, with a bachelor's degree, he is now the deputy general manager of smart city business department. His main research direction is traffic control and big data analysis, computer software development and application, system integration and security.

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

DESIGN OF INTERPLANETARY OBSERVATION TERMINAL BASED ON ALL PROGRAMMABLE SYSTEM-ON-CHIP

Haonan Jin[1], Lesheng He[1], Liang Dong[2],
Yongliang Tan[1] and Qingyang Kong[1]

[1]Information Institute, Yunnan University, Kunming, China
[2]Yunnan Astronomical Observatory, Chinese Academy of Sciences,
Kunming, China

ABSTRACT

The drastic changes in the solar wind will cause serious harm to human life. Monitoring interplanetary scintillation (IPS) can predict solar wind activity, thereby effectively reducing the harm caused by space weather. Aiming at the problem of the lack of the ability to observe IPS phenomenon of the 40-meter radio telescope at the Yunnan Astronomical Observatory of China in the frequency band around 300MHz, an IPS real-time acquisition and processing scheme based on all programmable system-on-chip(APSoC) was proposed. The system calculates the average power of 10ms IPS signal in PL-side and transmits it to the system memory through AXI4 bus. PS-side reads the data, takes logarithms, packages it, and finally transmits it to the LabVIEW host computer through gigabit Ethernet UDP mode for display and storage. Experimental tests show that the system functions correctly, and the PL-side power consumption is only 1.955 W, with a high time resolution of 10ms, and no data is lost in 24 hours of continuous observation, with good stability. The system has certain application value in IPS observation.

KEYWORDS

Interplanetary Scintillation, Solar Wind, All Programmable System-on-Chip, AXI4, LabVIEW.

1. INTRODUCTION

Interplanetary Scintillation (IPS) refers to the phenomenon that electromagnetic waves emitted from radio sources outside the solar system are scattered by the solar wind when they pass through the interplanetary space of the solar system, causing random fluctuations in electromagnetic intensity [1]. With the continuous development of human space exploration and aerospace technology, solar wind activity monitoring and space weather forecasting are becoming more and more important [2]. The drastic changes in the solar wind can cause magnetic storms, power transmission facilities to malfunction, interfere with ground shortwave communications, satellite communications, and even threaten the safety of spacecraft and astronauts [3]. Therefore, it is of great significance to predict solar wind activity by monitoring IPS signals.

At present, the common IPS monitoring methods are divided into direct measurement and ground-based measurement. Compared with direct measurement, ground-based measurement has a wider observation scale and is more economical and effective [4]. Ground-based measurement

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 75-86, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111906

modes are divided into single-station single-frequency (SSSF) and single-station dual-frequency (SSDF) and multi-station measurement. SSSF measurement mode obtains the scintillation power spectrum by means of spectrum fitting or characteristic frequency to obtain solar wind speed, such as Ooty Radio Telescope (ORT), located in India [5], and the 25-meter radio telescope in Urumqi, Xinjiang, China [6]. SSDF measurement calculates the solar wind speed by calculating the first zero frequency of the cross-correlation spectrum [7], which is more accurate than SSSF mode, such as Miyun 50-meter in Beijing, China [8]. The multi-station measurement mode can directly obtain the projected solar wind speed, such as STELab in Japan [9]. The 40-meter radio telescope at the Yunnan Astronomical Observatory in China is one of the important observation equipment for radio astronomy signals in China. Compared to Xinjiang 25-meter, Miyun 50-meter and FAST [10], Yunnan Astronomical Observatory 40-meter radio telescope is located in the south with low latitude, and can observe many radio sources that cannot be observed by other radio telescopes. According to the experience of scientists in IPS observations, when the observation frequency of radio telescopes is low, the radio source information is more abundant, such as 327MHz. Therefore, improving the receiving equipment of the 40-meter radio telescope near the 300MHz frequency band of Yunnan Observatory for IPS observation is of great significance to fill the gap in its IPS signal observation capability.

In view of the characteristics of the 40-meter radio telescope of the Yunnan Astronomical Observatory [11], the system designed in this paper is a SSSF mode. The characteristic time scale of the change of IPS signal received by a ground-based single station is from 0.1s to 10s [12]. In order to achieve long-term, stable and high-time resolution IPS signal monitoring, the system needs to have high-speed signal acquisition, high-speed cache, big data storage, high-speed processing and communication capabilities. Xilinx's proposed All Programmable System on Chip (APSoC) benefits from the high-speed interconnection between Programmable Logic (PL) and Processing System (PS) through the AXI4 bus [13]. Compared with the traditional FPGA+ARM/DSP processor architecture, the system has higher bandwidth and lower power consumption. Secondly, the data interface of PS-side is simple to implement and convenient to store. Through the cooperation of software and hardware, the system development cycle can be greatly shortened [14]. In summary, APSoC is very suitable for high-speed acquisition, processing and storage of IPS signals. Therefore, an APSoC based IPS observation scheme is proposed in this paper. The acceleration part of digital signal processing is constructed on FPGA (PL-side) to achieve the purpose of real-time IPS signal processing. A terminal parameter control program is constructed in ARM Cortex-A9 processor (PS-side) to rapidly control the device parameters of IPS observation terminal. The IPS observation data is transmitted to the host computer for display and storage through gigabit Ethernet UDP. Based on the above solution, a flexible, fast, and stable real-time observation terminal for IPS signals is realized through the cooperation of software and hardware, which provides a technical foundation for the Yunnan Astronomical Observatory to carry out subsequent IPS observations and has certain application value.

2. OVERALL SYSTEM DESIGN

2.1. Overall system structure

The system uses Digilent's Eclypse-Z7 hardware platform, and the processor is Xilinx's xc7z020clg484-1 all programmable system-on-chip, which is implemented based on 28nm Artix-7, and has 2 Cotex-A9 ARM processors, which can achieve excellent Performance to power ratio and maximum design flexibility [15], can meet the needs of real-time data acquisition and processing of IPS signals. The ADC data sampling board uses ZmodADC1410 with AD9648 as the core. The sampling frequency of ZmodADC1410 is 100MHz, and it can obtain two 14-bit

signals at the same time. The programmability and high resolution of this device make it a reliable choice for the receiving end of radio astronomy signals, and the program gain control. ZmodADC1410 and Eclypse-Z7 used in this system realize high-speed connection through SYZYGY standard interface. Compared with Pmod, SYZYGY standard interface has higher speed and bandwidth, and is smaller and cheaper than FMC interface. This paper adopts the system structure shown in Figure 1 to realize the IPS observation terminal.

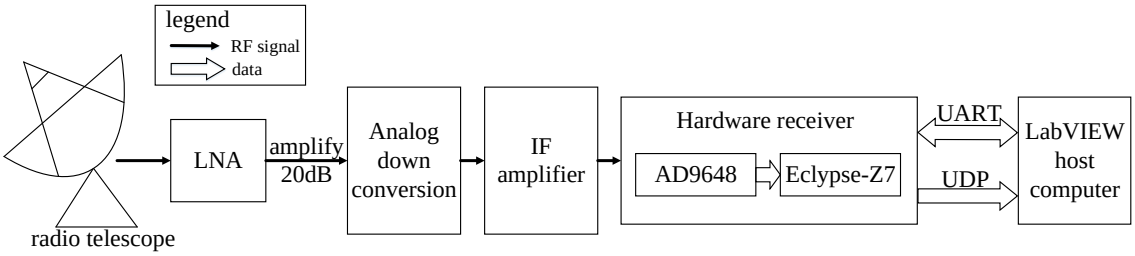


Figure 1. System overall structure block diagram

Firstly, the system uses a low noise amplifier (LNA) to amplify the weak radio signal received by the 40-meter radio telescope. Then, the analog downconversion module is used to downconvert the received 327MHz analog signal to obtain a 10MHz signal. Then, the intermediate frequency amplifier is used to amplify the intermediate frequency (IF) signal. The digital signal is input to PL-side of APSoC through AD9648. Real-time power calculation and superposition of the signal are carried out in PL, and then the superposition power data is averaged and stored in system memory through AXI4 bus for reading by PS-side. When the data is read, PS-side transmits the data to the LabVIEW host computer software in the form of Gigabit Ethernet UDP protocol for real-time display and storage. In addition, the system can also configure the gain, channel and coupling coefficient of the AD9648 in the hardware receiver through serial communication in the host computer application and issue control commands.

3. SYSTEM SOFTWARE AND HARDWARE DESIGN

In this paper, the main functions of APSoC are real-time processing and high-speed transmission of digitalized IPS signals, and flexible configuration of AD9648 parameters, including PS-side and PL-side. The PL-side is developed by Verilog with the HDL integrated development environment Vivado2019.1 provided by Xilinx company, and the PS-side is developed by C language in SDK2019.1.

3.1. PL-side system design

In order to achieve the high-speed acquisition and processing of IPS signals and ensure the stability of the system, the basic framework of IPS signal processing is constructed in the PL-side of APSoC. As the characteristic time scale of IPS signal changes is from 0.1s to 10s, in order to improve the system time resolution, the PL-side is used to calculate the power of signal sampling points within 10ms and calculate the average. Because the data interval between sampling points is very small, reaching nanosecond level, the process of power calculation and superposition needs hardware acceleration at PL-side. Then, the power data is mapped to the memory address through AXI4 bus for subsequent reading by PS-side.

According to the above scheme, this paper designs a Block Design project in Vivado. The overall block diagram of the system is shown in Figure 2. These include: IPS Data Processing module (Math IP), ZmodADC Control IP and AXI ZmodADC IP, Processor System Reset IP, AXI

Interconnect and ZYNQ core modules. In order to make the main signal easily visible, the ARESETN reset signal in the default figure is connected to Processor System Reset IP, and clock signal is also connected by default.

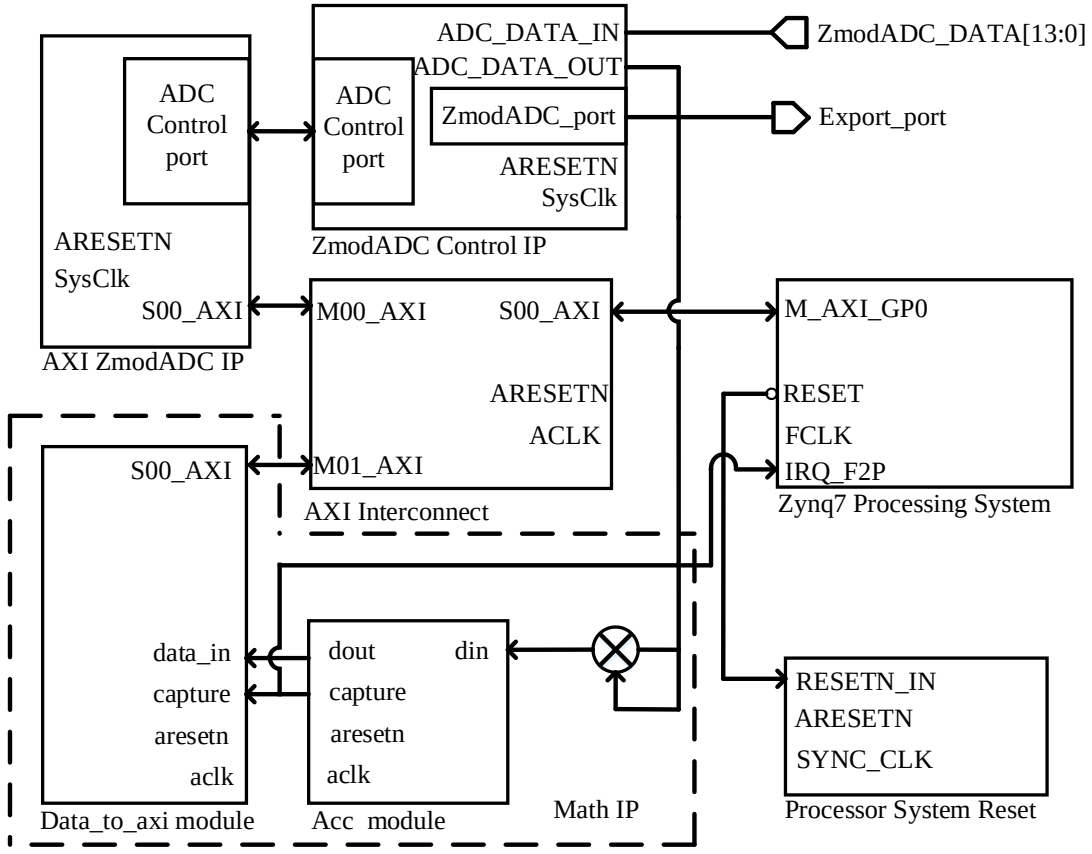


Figure 2. PL-side Block Desigen block diagram

The ZmodADC1410 Control IP core in Figure 2 is an ADC Control module, whose function is to realize data interaction between ZmodADC1410 and APSoC. PS-side can also indirectly Control the ZmodADC1410 Control IP by reading and writing the register of AXI ZmodADC IP via AXI4 bus. The two IPS are connected through a set of ADC Control ports, officially provided by Digilent. The ZYNQ7 Processing System is the hard IP address of the Cotex-A9 ARM processor. After the system was built, the global clock on PL-side was set to 100MHz in the ZYNQ IP core, that is, synchronous clock domain, in order to avoid the metastable problem caused by cross-clock domain. In addition, the serial port UART0, Ethernet port ENET0, GPIO, and the interrupt pin IRQ_F2P from PL to PS need to be enabled in the ZYNQ IP core for subsequent use. The AXI Interconnect IP realizes the topology of AXI bus, and connects M_AXI_GP0 of ZYNQ IP core to S00_AXI of AXI ZmodADC IP and Math IP, so as to realize the data communication between PL-side and PS-side.

The Data processing part (Math IP) consists of the power calculation module, power superposition module, and Data_to_axi IP to complete the calculation and superposition of the IPS signal power, and transmit the data to the PS-side for subsequent processing. Since dB is commonly used as a unit in engineering, logarithmic operation should be performed on the raw data. Since the power value to be sent is the average power value within 10ms, the amount of data is small, and a large number of LUTs will be consumed if the logarithm operation is

implemented using the PL-side. In order to save logical resources on PL-side, logarithm operation is carried out on PS-side. The schematic diagram of Math IP is shown in Figure 3. The detailed description of each part is as follows.

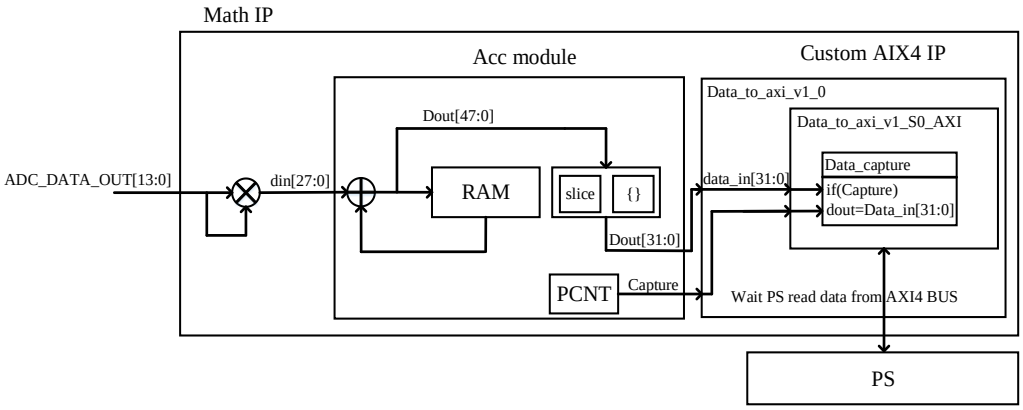


Figure 3. The schematic diagram of Math IP

3.1.1. Power calculation module

In order to calculate the IPS signal power, the amplitude of the sampling point needs to be squared to obtain the instantaneous power value of the point. The multiplication operation uses the Multiplier IP core. The data width of the ZmodADC Control IP output to the module is 14 bits. Therefore, the Multiplier IP input data width should be set to 14 bits for signed multiplication.

3.1.2. Power superposition average module

The function of this module is to calculate the average value of power superposition within 10ms. As shown in Figure 3, the superposition operation is realized by a RAM and adder. Set a PCNT counter, when the PCNT count reaches 1000000, Capture signal is raised, output result of the superposition named Dout[47:0], by intercepting data 28 bits higher to get the average signal power within 10ms. In order to adapt to AXI4 bus bit width, the power data is added to 32 bits and output to a custom AXI4 IP named Data_to_axi. In addition, the Capture signal also triggers an IRQ_F2P interrupt, which triggers a data read transaction on PS-side.

3.1.3. Data_to_axi module

In order to realize the data transmission between PL-side and PS-side, a custom AXI4 IP is created through the Tools option in Vivado, and add user logic. When the IP creation is completed, the IP is packaged for integration into Block Design. The IP hierarchy is shown in Figure 3. The Data_to_axi module internally instantiates a module named Data_capture. Whenever the Capture signal of the power superposition average module is raised, the 32-bit input data Data_in is loaded into the second read register of AXI4 IP at address 0x43C0008. The PS-side can read the value in the register by xil_In32() function reading the address.

3.2. PS-side program design

The PS-side clock frequency is 667MHz. In addition to logarithm taking and data packing operations, in order to achieve flexible parameter control, that is, to control the sampling channel, channel gain and coupling coefficient parameters of AD9648 through serial port, and to realize the functions of calling other IP cores and controlling gigabit Ethernet for data communication, it is necessary to build the PS-side operation program in the Cotex-A9 ARM processor. The main program flow of PS-side is shown in Figure 4.

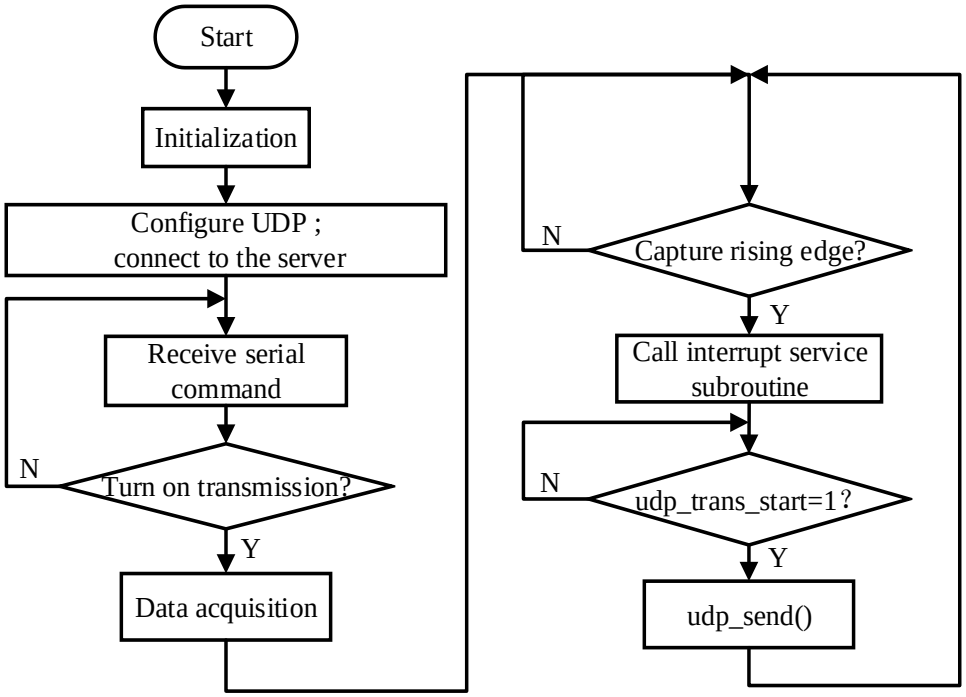


Figure 4. The main program flow chart of PS-side

Initialization includes the initialization of serial port, SPI, GPIO, ZmodADC1410 and LwIP (Light weight IP). The serial port uses uart0 of PS, and the baud rate of the serial port is 115200bps. ZmodADC1410 is initially set to channel 1, low gain mode, and the coupling coefficient is 0. Use the Lwip211 library integrated in SDK to program UDP communication.

After the initialization step is completed, configure the UDP related functions and connect to the server. When the serial port receives the transmission start command, enable AD9648. When Capture, the power data capture signal at the PL-side, is on the rising edge, the PL-to-PS interrupt is triggered. In the interrupt service subroutine, PS uses the xil_In32() function provided in xil_io.h file of SDK to read the power value stored in the system memory at the PL-side. Since the width of this data is 32 bits, after logarithmic operation, the width must be less than 16 bits. Construct data packets with the data type as uint32. Store the power data 16 bits lower and the data count 16 bits higher. After packing data packets, store them in udp_send_buffer to wait for sending, and the interrupt returns. In order to reduce the burden of data display and storage in the host computer, the sending buffer is set to send data once every 100 data is stored, When the sending buffer is full, udp_trans_start=1, the main program calls udp_send() function to transfer data to the LabVIEW host computer. In order to enable the host computer to distinguish one frame of data, set udp_send_buffer[0]=0xAABBCCDD, udp_send_buffer[101]=0xDDCCBBAA,

respectively, as the frame head and frame end. Since the data type of the buffer is uint32, a frame of data has a total of 408 bytes.

4. HOST COMPUTER APPLICATION DESIGN

In this paper, the visual programming software LabVIEW2018 of NI Company is used for the host computer application development. The main functions include:

(1) Serial port communication and UDP communication between the host computer and APSoC,
(2) Display and storage of IPS power data.

The host computer application adopts the string state machine program architecture, which is mainly composed of conditional structure and while loop. The program performs the corresponding state according to the string content by conditional structure. The UDP part and serial part of the application program of host computer are controlled by two state machines. The program flow chart of the LabVIEW host computer application is shown in Figure 5.

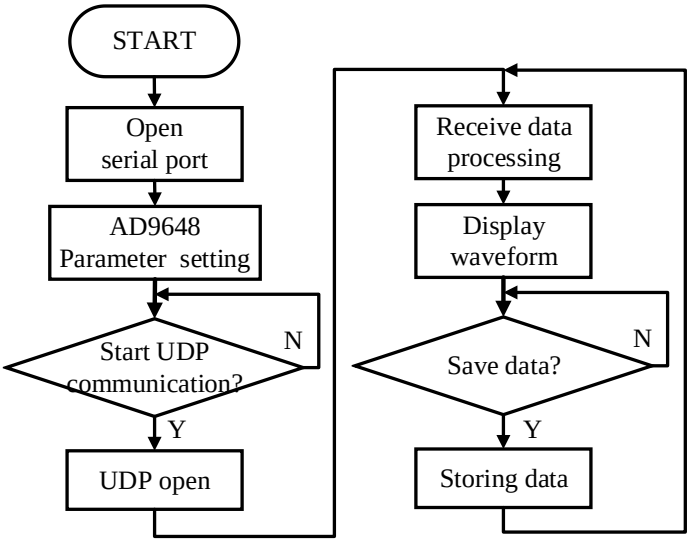


Figure 5. The program flow chart of the LabVIEW host computer application

First, Set and initialize the serial port, When the serial port runs normally, SERIAL_STATE indicator turns green. When the RX buffer of the serial port receives the response "Initialization successful" that APSoC is initialized, it can set the sampling channel, sampling length, channel gain, coupling coefficient and other parameters of AD9648 through the sending buffer according to the message prompted by the serial port. After the parameters of AD9648 are set, click SEND to send the parameters to APSoC. The PS-side of APSoC indirectly controls Zmod Control IP through AXI4 bus, so as to complete the flexible configuration of AD9648 parameters. After sending the start UDP transmission command, set the UDP target address and port, click to OPEN UDP CONNECT, UDP_STATE indicator turns green, you can see the IPS power signal sent from APSoC. Click SAVE DATA to obtain IPS power data saved in .jpg and .txt format in the specified path, which is convenient for subsequent analysis and use.

5. SYSTEM TEST AND RESULT ANALYSIS

5.1. Serial port and UDP function test

To verify the serial port and UDP functions, test according to the following steps:

Step 1: Set the IP address of the Eclypse-Z7 hardware platform to 192.168.1.150, port 8080, the IP address of the LabVIEW host computer to 192.168.1.151, port 8081, and connect the board and the host computer to the same local area network. Step 2: Open the LabVIEW host computer, select right COM number, set baud rate to 115200bps, open it. Step 3: Power on the board and keep the default AD9648 parameter. Step 4: RIGOL DSG815 signal generator is used to generate a signal with a power of -23dBm and a frequency of 327MHz, which is connected to the system. Step 5: Open the UDP connection and use the Wireshark to capture packets and compare the packets with the UDP receiving buffer of the host computer. The experimental results are shown in Figure 6 and Figure 7.

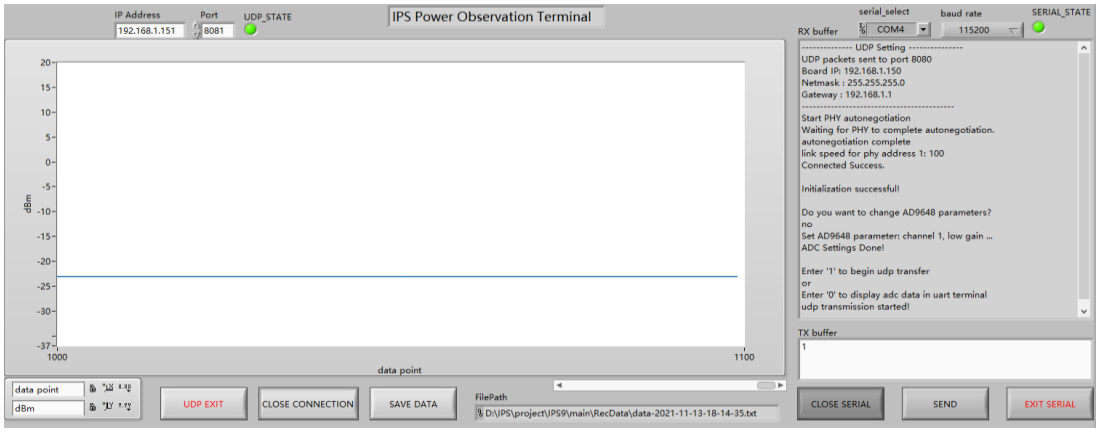


Figure 6. The LabVIEW host computer interface

As can be seen from Figure 6, at the moment when the board is powered on, the RX buffer of the serial port prints out “Initialization successful”, and sends the command, serial port also responds, which verifies that the serial port function is correct.

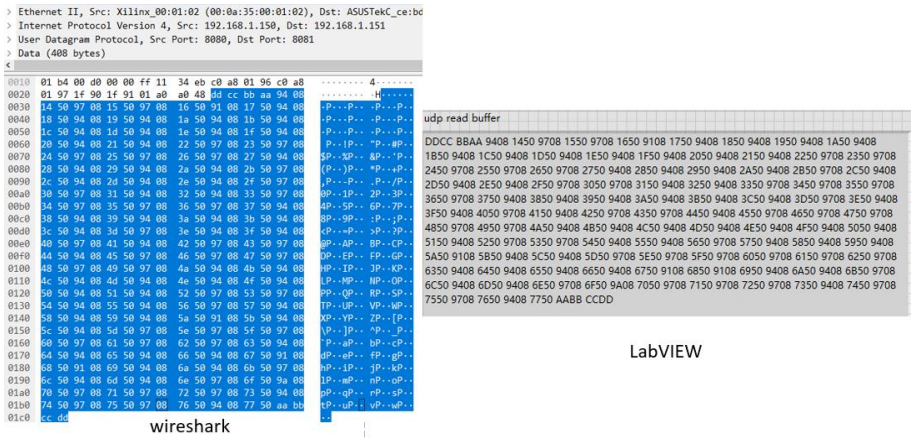


Figure 7. Comparison between Wireshark UDP packet capture and UDP read buffer of the LabVIEW host computer

As can be seen from Figure 7, the data in the Wireshark UDP packet capture and the LabVIEW host computer's UDP read buffer are exactly the same, indicating that the host computer successfully received the UDP packet from Eclypse-Z7, and the IPS signal power in Figure 6 is indeed -23dBm. It also verifies the correctness of the system function. The above experiments verify that the serial port and UDP functions of the host computer are correct and meet the design requirements.

5.2. Time resolution test

Since the characteristic time scale of IPS signal changes is 0.1s to 10s, in order to meet the requirements of the time resolution of the system in practical applications, it is necessary to test the time resolution of the system. Adjust the RIGOL DSG815 signal generator to a single step sweep mode, 30 points, 100ms dwell time, send a sine signal from -10dBm to -40dBm, and connect to the system for testing. Plot the output data of the system, as shown in Figure 8.

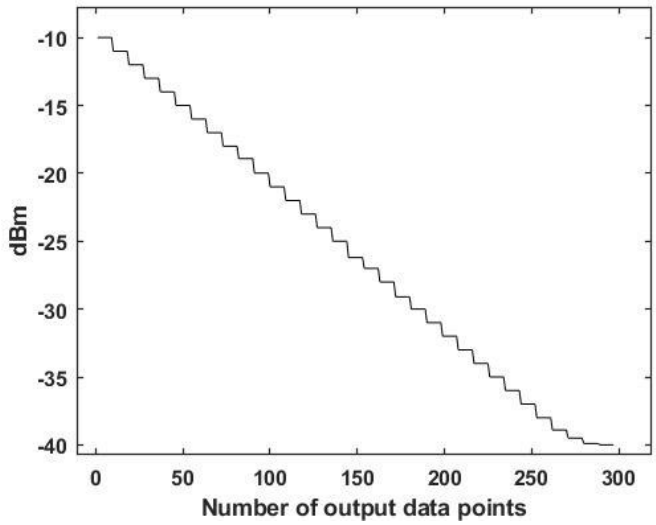


Figure 8. Single step scan mode system output data

As can be seen from Figure 8, when the signal power attenuates from -10dBm to -40dBm, the power value will change in real time with the change of the input signal. Secondly, when the time of a single step scan is 3s, the system output data points are 300 points, so the system time resolution can be calculated as 10ms. After repeated measurements using the above method, the system time resolution is 10ms, so it meets the design requirements.

5.3. Stability test

During the monitoring of IPS data, if the system has problems in the calculation, transmission, and storage of IPS power values, resulting in the omission of data points, the observation values of IPS data in this section will be meaningless and the precious observation time of radio telescopes will be wasted. In IPS observation, the longest observation time is only from sunrise to sunset [16], about 12 hours. In order to test the stability of the system, RIGOL DSG815 signal generator is used to send signals with power of -35dBm and frequency of 327MHz. Under the temperature condition of 20℃, continuous observation for 24 hours is carried out. Save system output data to a .txt file. According to statistics, there are 8640,000 data points in the final text file. The data waveform is drawn by MATLAB, as shown in Figure 9 (a). The first 1000 points

were intercepted and the fast Fourier transform was performed on the data within this time scale to obtain the IPS scintillation power spectrum with an integral time of 10s, as shown in Figure 9 (b).

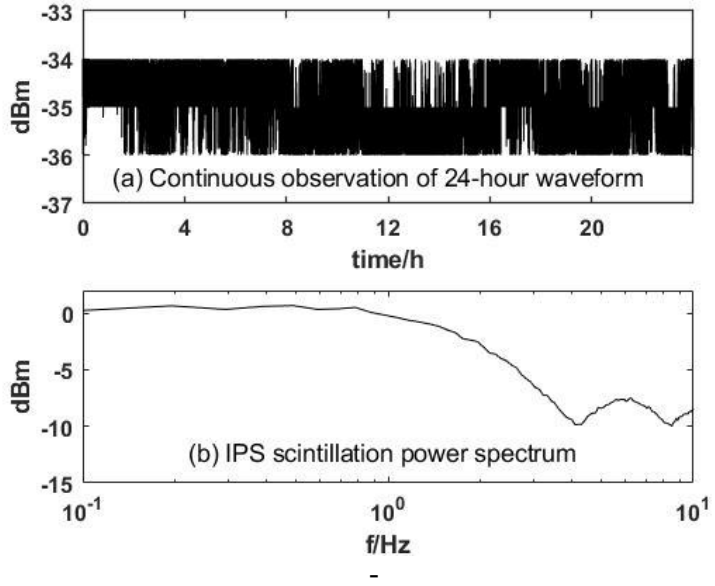


Figure 9. Continuous observation data and IPS scintillation power spectrum

As can be seen from Figure 9 (a), when the signal power is weak, the continuous observation results basically present a straight line. After calculation, the signal power fluctuation is about ±1dBm, indicating that the system has good test accuracy and stability, and meets the design requirements.

5.4. Programmable logic resource usage and power consumption

This system is based on Xilinx's XC7Z020CLG484-1 all programmable system-on-chip. In Vivado, after the PL part of the system is designed according to section 2.1, the final PL-side resource usage is shown in Table 1.

Table 1. PL-side resource usage

Resource	Utilization	Available	Utilization%
LUT	5552	53200	10.44
LUTRAM	513	17400	2.95
FF	8965	106400	8.43
BRAM	24.50	140	17.50
DSP	2	220	0.91
IO	32	200	16.00
BUFG	7	32	21.88

As can be seen from Table 1, BRAM and BUFG are the most used resources on PL-side of the system, accounting for 17.5% and 21.88% of the total resources respectively. DSP is the least used resource, and only 1% of the total resources are used, indicating that the system has a large amount of resource margin and good scalability. And the power analysis tool in Vivado was used to evaluate the power consumption of the PL-side of the system. The power consumption was 1.955W in actual operation, which met the design requirements of low power consumption of the system.

6. CONCLUSIONS

In this paper, we propose an IPS observation system based on Xilinx Zynq7000 APSoC. The system takes XC7Z020CLG484-1 as the core, realizes the real-time processing and transmission of IPS power signal by means of hardware and software cooperation, and realizes a good human-computer interaction interface based on LabVIEW. Experiments show that the system has high integration, high time resolution and good stability. The work in this paper has laid a technical foundation for the 40-meter radio telescope of Yunnan Astronomical Observatory to carry out IPS observation, and has certain application value. Next, we will consider integrating the post-processing of IPS signals into APSoC to improve system performance.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. U1631121). Thanks to all the contributors who have worked on this paper.

REFERENCES

[1] Liu, L. J. & Peng, B. , (2016). Overview of Interplanetary Scintillation Observation in my country. Science in China: Physics, Mechanics, Astronomy (6), 6.
[2] Yang Y. , Shen F. , Yang Z. , (2020). Three-dimensional interplanetary solar wind mhd simulation driven by various observation data. Chinese Journal of Space Science, 40(3), 10.
[3] Zezhou G. , Miao L. , Hongxiang W. , (2019). The threat of the sun-a sidelight from the discussion on the topic of "Solar Storm" at the Institute of Physics, Chinese Academy of Sciences. Physics (3), 2.
[4] Liu, L. J. & Peng, B. , (2010). Single-station single-frequency data processing of interplanetary scintillation. Astronomical Research and Technology (01), 25-30.
[5] O Be Roi, D. , & Rao, A. P. , (2000). Tomography of the solar wind using interplanetary scintillation. Journal of Astrophysics & Astronomy, 21(3-4), 445-446.
[6] Liu, L. J. , & Peng, B. , (2012). Observations of interplanetary scintillation in china. Proceedings of the International Astronomical Union, 8(S294).
[7] Liu, L. J. & Peng, B. , (2009). Single-station single-frequency and single-station dual-frequency simulation of interplanetary scintillation. Science in China: Series G (10), 6.
[8] Zhen X. , & Zhang. , (2007). A study on the technique of observing interplanetary scintillation with simultaneous dual-frequency measurements. Chinese Journal of Astronomy & Astrophysics.
[9] Kim, J. , & Jackson, B. V. , (2015). IPS Space Weather Research: Korea-Japan-UCSD.
[10] Liu, L. J. , Peng, B. , Yu, L. , Yu, Y. Z. , & Mejia-Ambriz, J. , (2021). A pilot study of interplanetary scintillation with fast. Monthly Notices of the Royal Astronomical Society, 504(4), 5437-5443.
[11] Hongbo Z. , Peifeng M. , Min W. , Jianjun Z. , Xinxin Z. , & Shuobiao S. , (2008). 40m radio telescope. Astronomical Research and Technology-Journal of the National Astronomical Observatory, 005(002), 187-191.
[12] Xuan Y. , (2020). Research on digital multi-beam synthesis technology based on interplanetary scintillation (IPS) telescope. (Doctoral dissertation, University of Electronic Science and Technology of China).
[13] He B. , & Yanhui Z. , (2016). Xilinx Zynq-7000 Embedded System Design and Implementation. Electronic Industry Press.
[14] Dalin Z. , (2019). Hardware design of Zynq-based software radio IF processing module. (Doctoral dissertation, University of Electronic Science and Technology of China).
[15] Ming Z. , Haibin W. , Yixin W. , Xinan Y. , & Xinbing C. , (2018). Design and implementation of pseudo-color image processing system based on zynq-7000. Electronic measurement technology.
[16] Liu, K. , Desvignes, G. , Cognard, I. , Stappers, B. W. , Verbiest, J. , & Lee, K. J. , et al, (2014). Measuring pulse times of arrival from broad-band pulsar observations. Monthly Notices of the Royal Astronomical Society(4), 3752-3760.

AUTHORS

Haonan Jin, Yunnan University, Main research on Astronomical signal processing and embedded development.

Lesheng He, Yunnan University, Associate Professor, Research on IoT security and weak signal processing.

Liang Dong, Yunnan Astronomical Observatory, Senior engineer, Research on Radio astronomy technology and its transformation, space weather.

Yongliang Tan, Yunnan University, Main research on digital signal processing and cryptography.

Qingyang Kong, Yunnan University, Main research on IoT security and cryptography.

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

DESIGN OF SRAM-BASED 8T-CELL FOR MEMORY ALIAS TABLE

Saleh Abdel-Hafeez[1, 2], Sanabel Otoom[1] and Muhannad Quwaider[1]

[1]Jordan University of Science and Technology, Dept. Computer Engineering, Irbid 22110, Jordan
[2]Sabbatical at Qassim University, Dept. of Computer Eng., College of Computer, Qassim, Buraydah, Saudi Arabia

ABSTRACT

Memory Alias Table exploits a major role in Register Renaming Unit (RRU) for maintaining the translation between logical registers to physical registers for the given instruction(s). This work presents the design of the memory Alias Table based on the 8T-Cell with multiport write, read, and content-addressable operation for 2-WAY three operands machine cycle. Results show that four read ports operate simultaneously within a half-cycle, while two-write ports operate simultaneously within the other half-cycle. The operation of content-addressable with two parallel ports is managed during the half-cycle of the read phase; thus, the three operations occur within a single cycle without latency. HSPICE simulations conduct 32-rows x 6-bit with 21T-Cell memory Alias Table that has 4-read ports, 2-write ports, and 2-content-addressable ports using a standard 65 nm/1V CMOS process. Simulations reveal that the proposed design operates within a one-cycle of 1 GHz consuming an average power of 0.87 mW.

KEYWORDS

Content-Addressable,8T-Cell SRAM, 2-WAY Instructions Cycle, Memory Alias Table, Register Renaming Unit

1. INTRODUCTION

The hardware of instruction-level parallelism constitutes several out-of-order units within the pipeline structure. These units facilitate the execution of multiple instructions within the stages of the pipeline to reduce the dependencies between instructions; and thus, improving the overall Amdahl's law measure performance [1]-[4]. One of these stages of the pipeline is the instruction dispatch unit that allocates mapping of the logical to the physical registers since the computations were held in physical implementations. Register Renaming Unit (RRU) holds this mapping as well as deallocates back the physical to the logical registers, wherein the number of the physical registers is greater than the number of the logical registers [5]-[7].

Register renaming unit constitutes primarily three tables with several comparators and priority encoders to orchestrates the mapping between logical and physical registers; and thus, increase the performance of out-of-order (OoO) speculative execution unit. The three tables are Alias Table, Physical Table, and Architectural Table [8]-[11]. The Alias Table is indexed by the logical register number and holds the mapping to the physical registers. The Physical Table shows the availability of each physical register. That is, if the allocated bit is set, the corresponding physical register is being mapped by a destination logical register of the assigned instruction;

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 87-106, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111907

otherwise, the corresponding physical register is free and can be allocated by a destination logical register of incoming instructions. The Architecture Table records all completed evaluated physical registers and their actual data.

Register renaming tables are considered high-cost design since each table requires several and multi-operations within a single cycle; besides, tables require to maintain low power consumption and short critical path for each operation [12]. Consequently, as the number of instructions per cycle increase and parallelisms becomes prominent in modern processors, the tables required to hold more parallel operations, and thus, having more complex circuits [13]. Therefore, the memory array of multiport SRAM cells considers an essential element in constituting the tables array, in which the cell structure provides several parallel operations with relatively reduce circuit complexity [14][15]. Still, comparators along with priority encoders in a form of prefix tree structure hold a large burden of design overhead area, wherein content-addressable memory (CAM) cell can provide a major role in reducing these complexities of design circuitry as well as maintaining the required operation efficiently [16]. Consequently, the Register Renaming tables have been investigated in several realizations of organizational memory structures. Some works focus on SRAM-based rather than CAM-based implementations for more scalable and energy-efficient at the cost of extra-overhead comparators and priority encoders [14][15]. Other recent works leverage the benefit of CAM to match the contents in a parallel fashion, such as hybrid SRAM-CAM structure [17][18] that result in reducing the latency cycles; and thus, improve throughput.

Another essential factor in the memory array is the type of cell structure. Some of the SRAM-base structures use the 6T-Cell [19], where the read and write ports share the same input-output bus; and thus, narrowing the noise margin. Subsequently, a highly sensitive sense amplifier with constant bias current is essential to boost the speed of the read operation. Other use the 9T-Cell structure [20], in which the back-to-back inverters of the write access port are disconnected by intermediate transistors to reduce the contention during write access mode. Thus, improving the power consumption for the write operation. However, this factor comes for the cost of double the size of the memory array as the authors mention due to the high cost of control logic that requires to generate the appropriate signal voltage level for the intermediate cell's transistor. Further attempts of SRAM-base cell is the use of 10T-Cell [21] and 7T-Cell [22] that are often used for ultra-low supply voltage in the order 0.5 V or even less for low power consumption with low operating frequency. A comparative study of the abovementioned memory cells exploits the characteristics of each cell and its effect on the overall memory array [23][24].

Nowadays, the 8T-Cell commonly use in most processors and graphics semiconductor companies in the field of CACHE and Shared Memory [25][26]. The 8T-Cell with sperate read and write ports structure considers an essential component for Register Renaming tables, wherein simultaneous read and write operations at different rows are needed. Therefore, multiport for write and read can realize fast access time with a full rail noise margin and low power consumptions [27][28]. Additionally, the 8T-Cell can be adapted to verities and a wide scale of semiconductor technologies, while maintaining its low-power and high-speed features [29]. Therefore, the 8T-Cell is considered in this work for constituting the Register Renaming tables. The 8T-Cell is reconfigured for multiport read and multiport write; besides, the CAM circuitry has added to the 8T-Cell forming a hybrid SRAM-CAM cell. Further facilitation of 8T-Cell reduces design time to market by introducing an automation algorithm from schematic layout to physical layout with the support of 8T-Cell library standard cell components [30][31].

As a result, the prior work on the Register Renaming unit architectural development [32] is extended to pay attention to the Alias Table circuit development as the most complex circuitry, which exploits several simultaneous operations within a single cycle [33]. The Alias Table

constitutes a memory array of SRAM-CAM cells, where the SRAM is based on the 8T-Cell and the CAM is based on the two pass-gate transistors forming XOR logic connected to another two pass-gate transistors of the match line. The SRAM part of the cell has four read ports and two write ports, while the CAM part has two content-addressable ports, results in a total of twenty-one transistors cell. The match line is pre-charged to the power supply, where any mismatch between cell content and the coming data from the mask register drops the match line voltage to the ground. Thus, if all cells in a particular row match the data of the mask register, the match line for that particular row preserves the power supply pre-charge value, which enables the Tri-state buffer to release the row index to the output match port. In summary, the proposed Alias Table circuit has the following key features:

- The Alias Table leverages the 8T-Cell SRAM structure that is suitable for continued low-cost CMOS technology with high-speed and low-power operations.
- The Alias Table conducts write, read, and content-addressable match index within a single clock cycle.
- The Alias Table provides four parallel read operations.
- The Alias Table provides two parallel write operations.
- The Alias Table provides two parallel content-addressable operations.
- The content-addressable operation compares all rows in parallel with the mask register and releases the associated match index address.

The remainder of this paper is organized as follows. Section II discusses the design of the SRAM-CAM cell circuit structure. Section III realizes the overall Alias Table circuit architecture with the proposed memory array of SRAM-CAM cells. Section IV gives the estimated overall critical path delay for each operation, while section V provides the HSPICE simulations and verifications. Additionally, section VI illustrates some performance features along with some comparison of recent works. The conclusion is given in Section VII.

2. SRAM-CAM CELL CIRCUIT

The cell given in Figure 1 is designed based on the well-known 8T-Cell circuit with the addition of a new CAM structure. The cell combines three circuits operations that are - four parallel read ports, two parallel write ports, and two parallel content-addressable ports. All three operations are independent of each other's and have separate ports, given a rail-to-rail noise margin operation that is attractive for continued technology scaling with low power supply voltage and high-speed operation. Moreover, the ports within the same operation are activated in parallel since they are gated with separate pass-gate transistors. TABLE 1 depicts the input/output signals' abbreviations and descriptions for the 21T-Cell.

The two write ports have a similar circuit structure to the 8T-Cell write port, where the back-to-back inverters are flipped by input bit and inverted input bit. Subsequently, two input bits are gated by two independent pass-gate transistors from each side of the back-to-back inverters given the write circuit with eight transistors count. Each write port is associated with two pass-gate transistors that were gated by a write decoder. Therefore, the two separate write decoders activate the two separate write ports, where each port has a separate data bit.

TABLE 1. 21T-Cell signals definitions and abbreviations

Input-Output Signals	Representations
DA	First Input
DA!	First Inverted Input
DB	Second Input
DB!	Second Inverted Input
WDA	First Write Word Line Decoder
WDB	Second Write Word Line Decoder
RDA	First Read Word Line Decoder
RDB	Second Read Word Line Decoder
RDC	Third Read Word Line Decoder
RDD	Fourth Read Word Line Decoder
OA	First Output
OB	Second Output
OC	Third Output
OD	Fourth Output
MLA	First Match Line
MLB	Second Match Line
MA	First Mask Register
MA!	First Inverted Mask Register
MB	Second Mask Register
MB!	Second Inverted Mask Register
CLK	System Clock

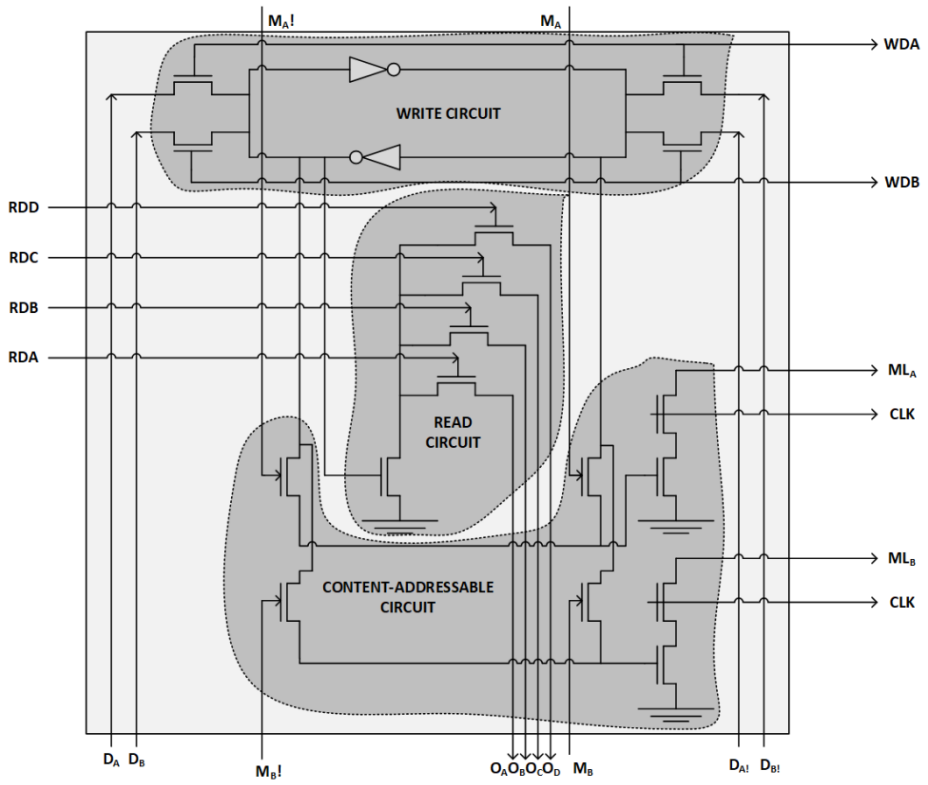


Figure 1. The circuit structure of the proposed SRAM-CAM 21T-Cell

Subsequently, two different data bits can be written to two different cells simultaneously of the same column of the memory array. A priority encoder is realized at each two input data bits in order to assure no contention of bits values is stored simultaneously at the same cell. Thus, precludes any possibility of two different data bits to be written simultaneously at the same cell.

The four read ports exploit the same structure of the 8T-Cell read port, and they can operate in parallel since each port has a separate pass-gate transistor gated with a separate decoder. Subsequently, the read circuit has five transistors. The discharge-transistor for all the four pass-gate transistors is designed with a larger width ratio than in the 8T-Cell single read port since it has more diffusion capacitances at its drain node. The four output ports are pre-charged to the power supply voltage through the output buffers. Once the read ports' decoders are enabled, the output ports are either discharged to the ground or preserved the pre-charged voltage, based on the cell stored value.

The CAM circuit compares the cell value with an inverted mask bit value and vice versa by using two pass-gate transistors with a common drain forming an XOR logic structure. Additionally, the second content-addressable port has a similar structure, but the cell value is compared with the second mask bit that is related to the second mask register. For brevity of discussion, the detailed operation of one port content-addressable is presented since the second port has similar behavior. The common drain out of the XOR logic is directed to the gate of discharge pass-gate transistors, which is connected with the match line through a gated clock pass-gate transistor. During the low

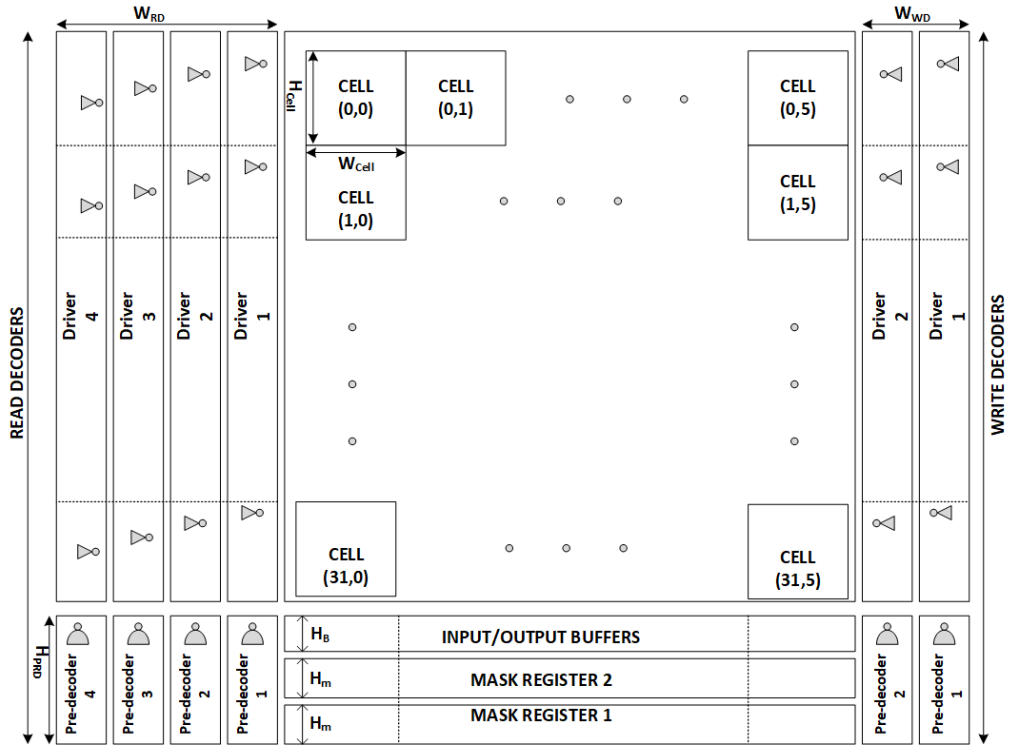


Figure 2. The architectural diagram of the proposed Alias Table

Phase of the clock, the match line is pre-charged to the power supply voltage. Consequently, during the high phase of the clock, the match line is either discharged to the ground or preserve the pre-charge value.

As a result, the cell contains two write ports, four read ports, and two content-addressable ports with twenty-one transistors count. The cell used similar geometry sizes (L/W) for the well-known 8T-Cell for write and read operations with similar layout features of combining three metals materials that detail some highlights on the cell layout structure and geometry sizes [34][35].

3. ALIAS TABLE CIRCUIT ARCHITECTURE

The study for the Register Renaming unit shows a machine with two simultaneous fetch instructions, where each instruction has three operands, describes the requirements for Alias Table design. The proposed Alias Table constitutes the memory array of size 32-row X 6 bit. Such that, the 32-row presents the logical indices, while the 6-bit presents the physical indices. Some machines present 7-bit instead of 6-bit per each row in order to allocate the last bit for availability [16][18]. However, in this context,the 6-bit is used for the brevity of discussion and ease of understanding. Each cell in the memory array is a hybrid SRAM-CAM of 21T-Cell describes in section II since the Alias Table would require four simultaneous reads, two simultaneous writes, and two simultaneous contents-addressable. An additional constraint is required to maintain all three operations within one cycle of the pipeline stage. Figure 2 demonstrates the topology block diagram of the overall Alias Table realizes four read decoders, two write decoders, a memory array of 21T-Cell, two mask registers, and input-output buffers with priority encoders.

As clearly illustrated in Figure 2, depicting the geometry sizes of the 21T-Cell, which is the height (H_{Cell}) and the width (W_{Cell}), the memory array can be estimated as the height of 32 x H_{Cell} and the width as 6 x W_{Cell}. The four read decoders' drivers are aligned with the pitch size of the cell height (H_{Cell}), where the width of the read decoders (W_{RD}) can simply be estimated. Similarly, the two write decoders' driver width (W_{WD}) is depicted.

On the other hand, each bit width of the mask register is aligned with the pitch cell width (W_{Cell}), while the mask bit height (H_M) is estimated. Similarly, each input/output buffer is aligned with the pitch cell width and the buffer height (H_B) is estimated. The read pre-decoders height (H_{PRD}) is aligned with the height of the input/output bus of the array (H_M+H_B), where the width of the read pre-decoder is aligned with the decoders width (W_{RD}) as clearly shown in Figure 2. Similarly, the write pre-decoder width is aligned with the write decoder width (W_{WD}), and the height of the write pre-decoder is aligned with the input/output bus of the array (H_M+H_B). As a result, the total layout geometry of the Alias Table circuitry can be estimated and measure, as well be depicted in simulation section V for the given technology parameters.

4. CIRCUIT AND TIMING ANALYSIS

The circuit detail of each operation is considered in the following subsections along with some highlights on the time delay model for each operation by considering only the critical path. Subsequently, the acronym "TD_{NAME}" is referred to the time delay, where the subscript "NAME" is referred to the circuit's component involved in the critical path. Furthermore, the scalability of the approach is evaluated using the timing of all critical paths concerning a one-unit gate delay (GD), which provides an analysis that is independent of technology factors for direct comparison purposes [36]. In simulation section V, the 65 nm/1 V technology parameters and HSPICE simulator are used to cross-verify the derived critical path for each operation. The proposed design uses only basic CMOS logic gates structure with basic width/length sizes to provide design layout clarity and cost-effectiveness for continued technology scaling.

4.1. Read Circuit Architecture and Timing

The Alias Table shown in Figure 3 presents the read portion circuit architecture. Each cell in the memory array has four read ports that can be accessed in parallel by the four read decoders due to a separate pass-gate transistor on each port. Therefore, all the cells within a column share the same four read ports lines; thus, each column has a bus of four output lines. Each output line has an output buffer that has a pre-charge PMOS transistor, given a total of twenty-four output bus lines. During the low phase of a clock, all the twenty-four output bus lines are pre-charged by PMOS transistors to a supply voltage, which is known as a pre-charge phase. Inversely, all the twenty-four output bus lines are disabled from pre-charging during the high phase of a clock, which is known as an evaluate-phase. During the evaluation phase, each read decoder enables a one-row port of the memory array, and thus, the four decoders enable four-row ports of the memory array. The four-row ports can be a particular row of the memory array or separate rows of the memory array since the read decoders along with the read ports are independent of each other.

Therefore, the read access time is started by the rising edge of the clock (CLK), such that, the total read access time for fetching the data from a particular row's cells to output bus bits is estimated
as follow:

$$TD_{Racc} = TD_{Decoder} + TD_{Row} + TD_{Cell\text{-}line} + TD_{Buffer} \quad (1)$$

That is, the TDDecoder is the decoder time delay since all four decoders are activated in parallel and each decoder has a separate 5-bit input address bus. Each read decoder has a simple structure of pre-decoders and simple inverter drivers, in which the pre-decoder has a simple prefix-tree of NAND-Gates where the last gate is gated with the clock. Thus, the decoder time delay is

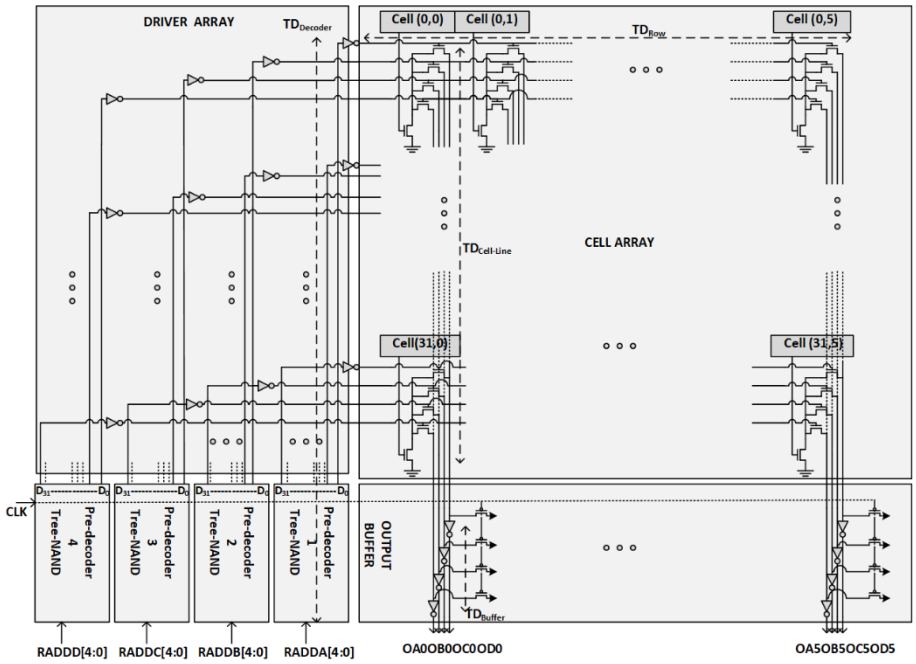


Figure 3. The proposed four parallel ports read circuitry with critical path

approximated into $TD_{Decoder}$ = 5 GDs to consider the worst-case scenario; wherein the decoder delay is usually designed to be even less than 5 GDs for a similar memory size [34]. Furthermore, each read word line is driven by a simple inverter from the decoders' drivers' circuit, and each read wordline has a parasitic of six capacitive gates as clearly shown in Figure 3. Consequently, the read wordline time delay can be approximated into TD_{Row} = 3 GDs as a worst-case scenario.

Once the four read word lines are activated, the data is fetched into the output lines; and thus, the worst-case scenario is to propagate the data to the output lines passing through thirty-two rows, where each row has one parasitic capacitance of type depletion region of NMOS transistor as illustrated in Figure 3. Thus, each bit line of the output bus has a total of thirty-two type parasitic capacitances of the depletion of NMOS transistors. HSPICE simulations shows the depth can be approximated with $TD_{Cell-line}$ = 8 GDs as a worst-case scenario. The bit line output buffer constitutesa simple inverter with a simple pre-charge PMOS transistor, or more often a simple latch to holdthe data for a complete clock cycle for other interfacing components. Subsequently, the latch accesstime is only TD_{Buffer} = 1 GD. As a result, the total read access time estimated by Eq. (1) is:

TDRacc = 5 GDs + 3 GDs + 8 GDs + 1 GDs = 17 GDs.

4.2. Write Circuit Architecture and Timing

The write circuit shown in Figure 4 exploits the Alias Table for the write operation. The detailed of the write portion of the cell in the memory array is addressed, which contains two parallel

write ports that are activated by two separate write decoders. Therefore, each column has two input data bit lines, giving a total input data bus of twelve-bit lines. A priority encoder depicted in Figure 5 streamlines the input data at each column to prevent different data written to the same cell; however,

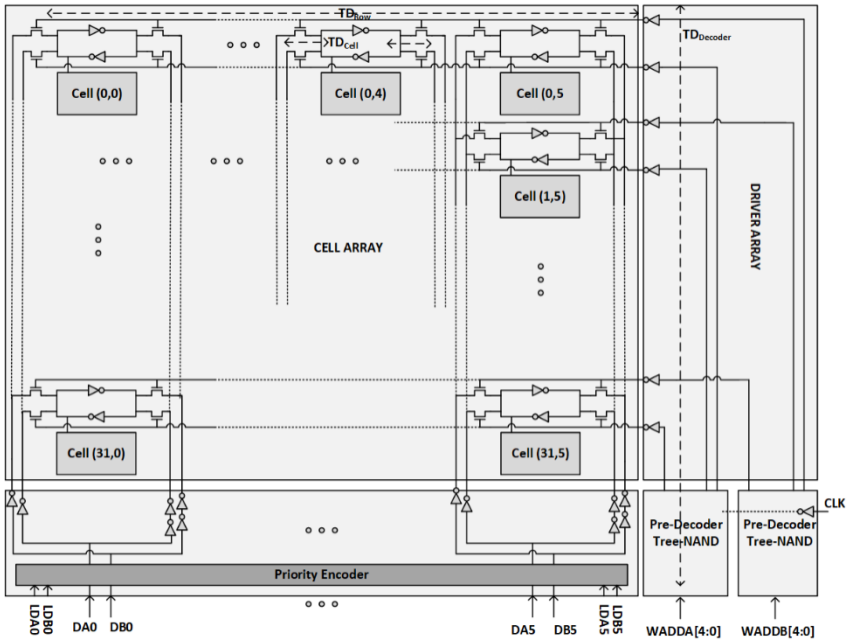


Figure 4. The proposed two parallel ports write circuitry with critical path

the priority encoder permits different data written in different cells. Thus, a total of six priority encoders exploits the input data buffer to leverage the advantages of two write operations to two different rows simultaneously and prevent the contention of different data to the same row.

The write operation begins during the low phase of the clock, where the access time can simply be derived by the total number of GDs as follow:

$$TD_{Wacc} = TD_{Decoder} + TD_{Row} + TD_{Cell} \quad (2)$$

That is, the $TD_{Decoder}$ is the decoder delay time, which is similar to the read decoder structure. Thus, the write decoder delay time is $TD_{Decoder}$ = 5 GDs as a worst-case scenario. Furthermore, each writes word line is associated with parasitic capacitances of type capacitive NMOS gates of count twelve since there are six columns of the memory array as clearly shown in Figure 4. Consequently, the write word line time delay can be approximated into TD_{Row} = 6 GDs as a worst-case scenario. Moreover, each bit of input data is propagated through the column of data bit lines waiting for the particular rows to be enabled by the write decoder, where the propagation time occurs in parallel with the decoder time. Therefore, the write delay cell time is only the count time to flip the cells of the row, which is approximated as TD_{Cell} = 1 GD. As a result, the write access time using Eq. (2) is:

TD_{Wacc} = 5 GDs + 6 GDs + 1 GDs = 12 GDs.

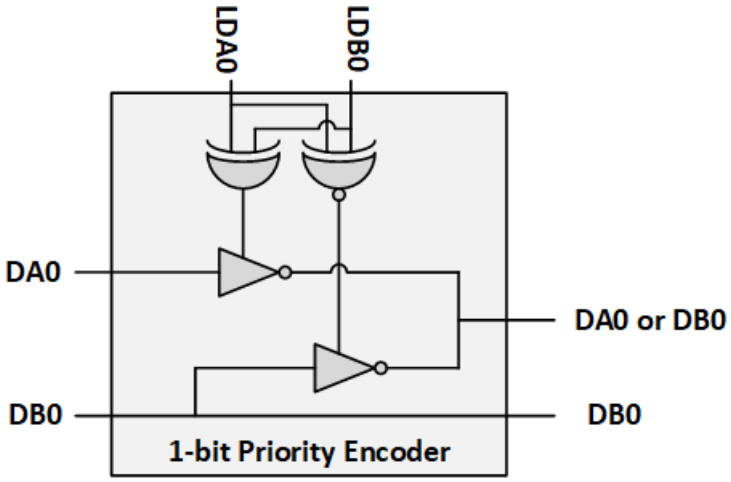


Figure 5. The proposed Priority Encoder circuitry

4.3. Content-Addressable Circuit Architecture and Timing

The Alias Table in Figure 6 illustrates the content-addressable circuit operations, which constitutes the CAM portion of the cell in the memory array, the match lines, the two mask registers, and the Tri-state output buffers. Two match lines per row are supposed to be compared simultaneously with all bits of the mask registers; however, only one match line per row is presented for the brevity of discussion. Each match line on each row is pre-charged to a power supply by a PMOS transistor during the low phase of the clock. During the high phase of the clock, each bit of the mask register is broadcasted to all CAM cells in the associated column of the memory array by two mask lines. Subsequently, each CAM cell in the column examines the mask bit against its content, and thus, the CAM discharges or preserves the pre-charge voltage of the associated match line based on the comparison with the cell's content. If all CAM cells in the row hold the match line voltage; then, the mask register matches the content of that particular row cells. Thus, the match line of that particular row enables the associated memory array index through a Tri-state buffer as illustrated in Figure 6. This matched index propagates to the match-output port of the Alias Table.

The content-addressable of the Alias Table evaluates the match address index at the high phase of the clock, giving the access time as follow:

$$TD_{Cacc} = TD_{FF} + TD_{Mb\text{-}line} + TD_{XOR} + TD_{Match\text{-}line} + TD_{Tri\text{-}state} + TD_{Buff\text{-}out} \quad (3)$$

That is, the TD_{FF} is the D-Type Flip-Flop access time of mask register, TD_{XOR} is the cell's pass-gate access time, $TD_{Tri\text{-}state}$ is the Tri-state buffer access time, and $TD_{Buff\text{-}out}$ is the output buffer access time, which all are more-less have the same access time delay TD = 1 GDs. The delay of the mask bit lines propagated through each column, and thus, passing through thirty-two cells of the parasitic type NMOS gate, in which two parasitic NMOS gate per cell since there are two match lines per cell. Consequently, each match bit lines heavily inherited with sixty-four parasitic gate capacitances. However, the mask bit lines are driven from the mask register, which has an output buffer on each of the mask bit line. Therefore, the estimated delay time for the mask bit line is $TD_{MB\text{-}line}$ = 8 GDs. On the other hand, each match line has six parasitic capacitances of type depletion NMOS transistors, where the estimated delay is $TD_{Match\text{-}line}$ = 6 GDs since the worst-case delay is to discharge the match line by only one cell through two in

series pass-gate NMOS transistors. As a result, the content-addressable access time delay is derived from Eq. (3) as:

TD_{Cacc} = 1 GD + 6 GD + 1 GDs + 8 GDs + 1 GDs + 1 GDs = 18 GDs.

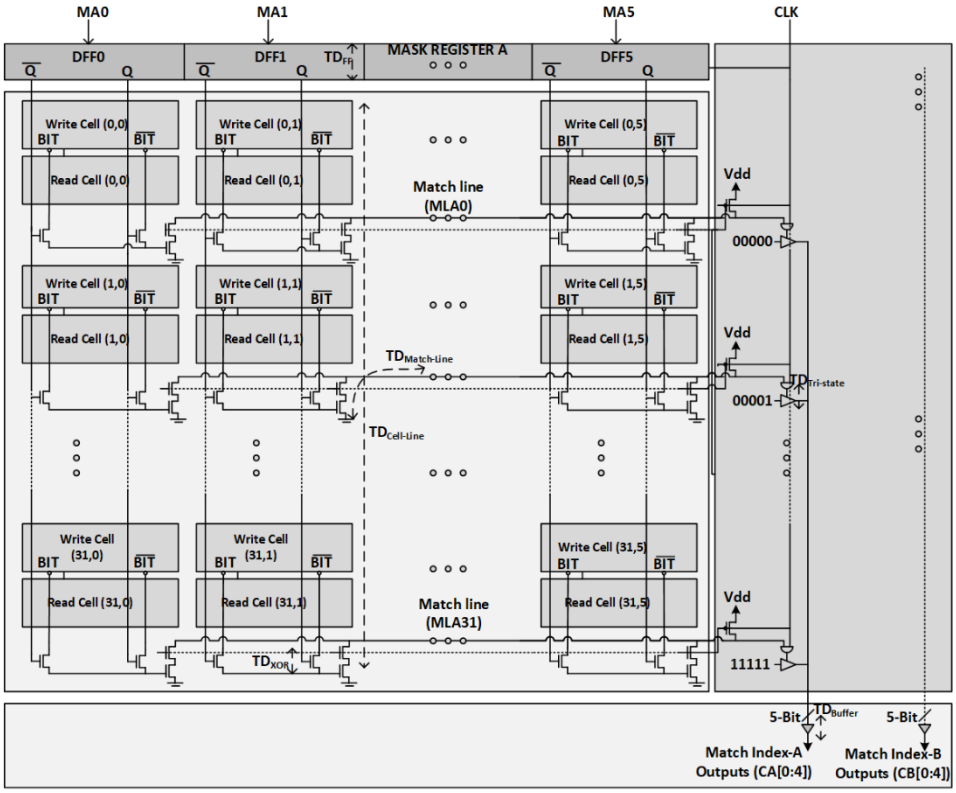


Figure 6: The proposed two parallel ports content-addressable circuitry with critical path. Second port circuitry is omitted for ease of understanding, as described.

5. HSPICE SIMULATIONS

The memory Alias Table with the derived SRAM-CAM cell of 21 transistors is designed and tested of memory array size 32-row x 6-bit for two write ports, four read ports, and two content-addressable ports. Based on the proposed design architecture, the read, and the content-addressable occur during the high phase of the clock, while the write occurs during the low phase of the clock, given all three operations within a single clock cycle. All timing delay values, total power consumption, and total transistor counts are collected based on the cost-effective CMOS transistor level of 65-nm Taiwan Semiconductor Manufacturing Company (TSMC) technology with a 1 V power supply [37] using an HSPICE simulator [38]. Each standard GD estimates at 0.005 ns, but the delay model assumes GD = 0.02 ns as a precaution measure. Although all distributed fan-in and fan-out logic circuits are composed with a four-gate tree structure and minimum geometry (W/L) sizes.

The HSPICE simulations of memory Alias Table realizes two simultaneous write operations, four simultaneous read operations, and two simultaneous content-addressable operations, where all operations occur within a single cycle of the clock. Exhaustive simulations verify the corner cells of the memory array, which are the upper right-hand corner, the upper left-hand corner, the lower right-hand corner, and the lower left-hand corner. Besides, verifying setup and hold time

between signals and clock as well as the pre-charge time. The parasitic model for all signals' wires is estimated based on three layers of metals and 65-nm TSMC technology. Further simulations detail can be found in [34][35]. However, for the brevity of discussion and ease of understanding, the worst-case time delay is shown for each operation as only one simulation. Wherein, the reader can easily follow and verify.

5.1. Read Simulations

The simulation in Figure 7 is conducted for the design read circuitry of the Alias Table, which realizes in the circuit diagram of Figure 3. Assume address "0" assigns to read decoders "A" and "B", while address "31" assigns to read decoders "C" and "B". Since there is only six output bus due to six column and every bus has four read ports, the release output ports from the memory array row "0" are- "OA0, OB0, OA1, OB1, OA2, OB2, OA3, OB3, OA4, OB4, OA5, OB5"; additionally, the release output ports from row "31" are-"OC0, OD0, OC1, OD1, OC2, OD2, OC3, OD3, OC4, OD4, OC5, OD5". Figure 7 presents only the last read ports of corner cell row "0",
which is "OA5, OB5", and row "31", which is "OC5, OD5", instead of showing the complete ports of rows "0" and "31". Subsequently, only the read ports "A and B" of Cell(0,5) and "C and D" of Cell(31,5) are presented for the last column of the memory array as a worst-case timing scenario.

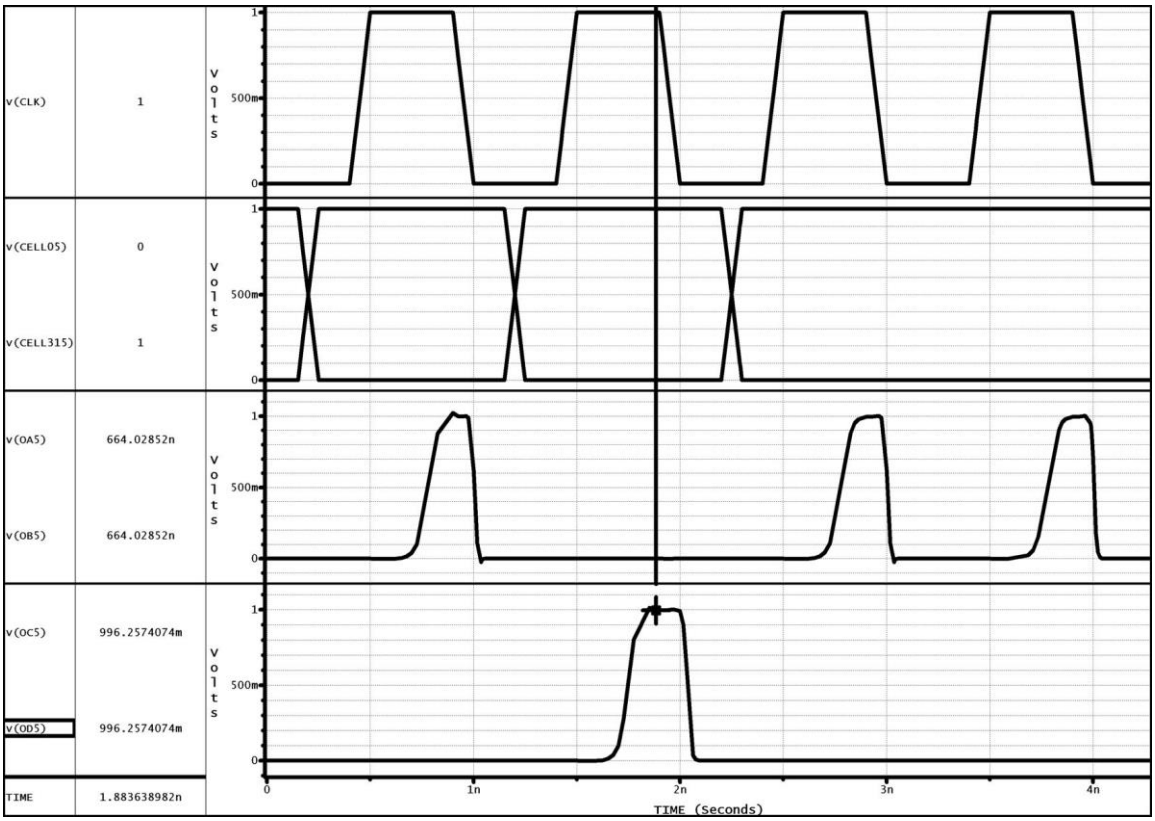


Figure 7. HSPICE simulation for four parallel ports read operation. Two ports fetch from row "0"and two ports fetch from row "31", where only last column is presented, as described.

The first row of Figure 7 gives the system clock running at 1 GHz, where the second row shows the data stored in Cell(0,5) and Cell(31,5), which are opposite to each other to give more simulation clarity to the reader. The third row shows the outputs "OA5, OB5" at about 0.33 ns

from the clock rising edge. These outputs present the fetch data value from Cell(0,5) using the decoders A, B that select row ”0” (read world line "0" is not shown in the simulation). The fourth row shows the outputs “OC5, OD5” at about 0.3 ns, which present the fetch data value from Cell(31,5) since the decoders C, D were selecting row ”31” (read world line "31" is not shown in the simulation). Notice, when the clock is asserted low the output bus shows a “0” value since all read lines hold the pre-charge voltage “VDD”, which were inverted at the read output buffer bus. The read output bus shows the data is fetched out at less than 0.35 ns that is very close to the derived read cycle time in Eq. (1), which is:

Read Access-Time = 0.02 ns/GD * 17 GDs = 0.34 ns.

As a result, the estimated delay model has more conservative results than the simulation model since the delay model scenario assumes more parasitic to further adhere to the safety margin delay. In either case, the read operation can safely operate at the half cycle of 1 GHz since the other half cycle holds the write cycle in parallel with pre-charge time.

5.2. Write Simulations

The write simulation in Figure 8 is conducted for the design write circuitry of the Alias Table, which realizes in Figure 4. Assume address "0" assigns to write decoder "A" and address "31" assigns to write decoder "B". Thus, the write world line "0" of port "A" (WLA0) and the write world line "31" of port "B" (WLB31) are enabled. The data input ports of the memory array are- “DA0, DB0, DA1, DB1, DA2, DB2, DA3, DB3, DA4, DB4, DA5, DB5”, which propagate through all the columns of the memory array to store on rows "0" and "31". Such that, data of port "A" stored on row "0" cells, while data of port "B" stored on row "31" cells.

Figure 8 presents the write simulation of Cell(0,5) and Cell(31,5) of the last column of the memory array as a worst-case timing scenario. The first row of Figure 8 gives the system clock running at 1 GHz. The second row shows the write word lines of rows "0" and "31" (WLA0, WLB31), which are asserted at the falling edge of the clock and de-asserted at the rising edge of the clock consuming a delay of about 0.12 ns. The third row shows the data input ports “DA5, DB5”, which propagates through the last column of the memory array scanning for the active word lines WLA0 and WLB5. The last two signals ("Cell05" and "C315") show the stored data of Cell05 using data port "A" and Cell315 using data port "B". The write simulations show the data is stored at less than 0.18 ns, result in a close write access time to the derived write cycle time in Eq. (2), which is:

Write Access-Time = 0.02 ns/GD * 10 GDs = 0.2 ns.

Consequently, the write operation is active at the low half cycle of 1 GHz, while the read operation is active at the high half cycle of 1 GHz. The pre-charge mode for the read and the content-addressable is active during the low half cycle of 1 GHz.

5.3. Content-Addressable Simulations

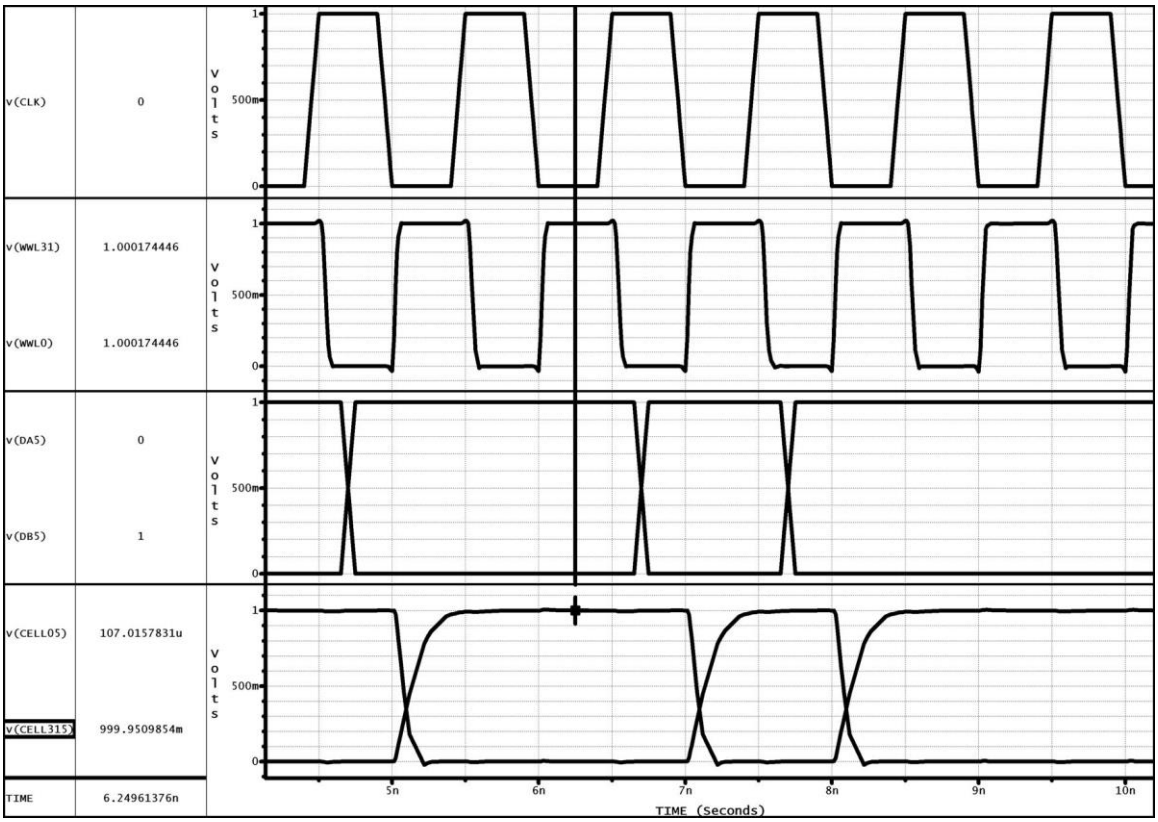


Figure 8. HSPICE simulation for two parallel ports write operation. Two ports fetch from row “0” and two ports fetch from row “31”, where only last column is presented, as described.

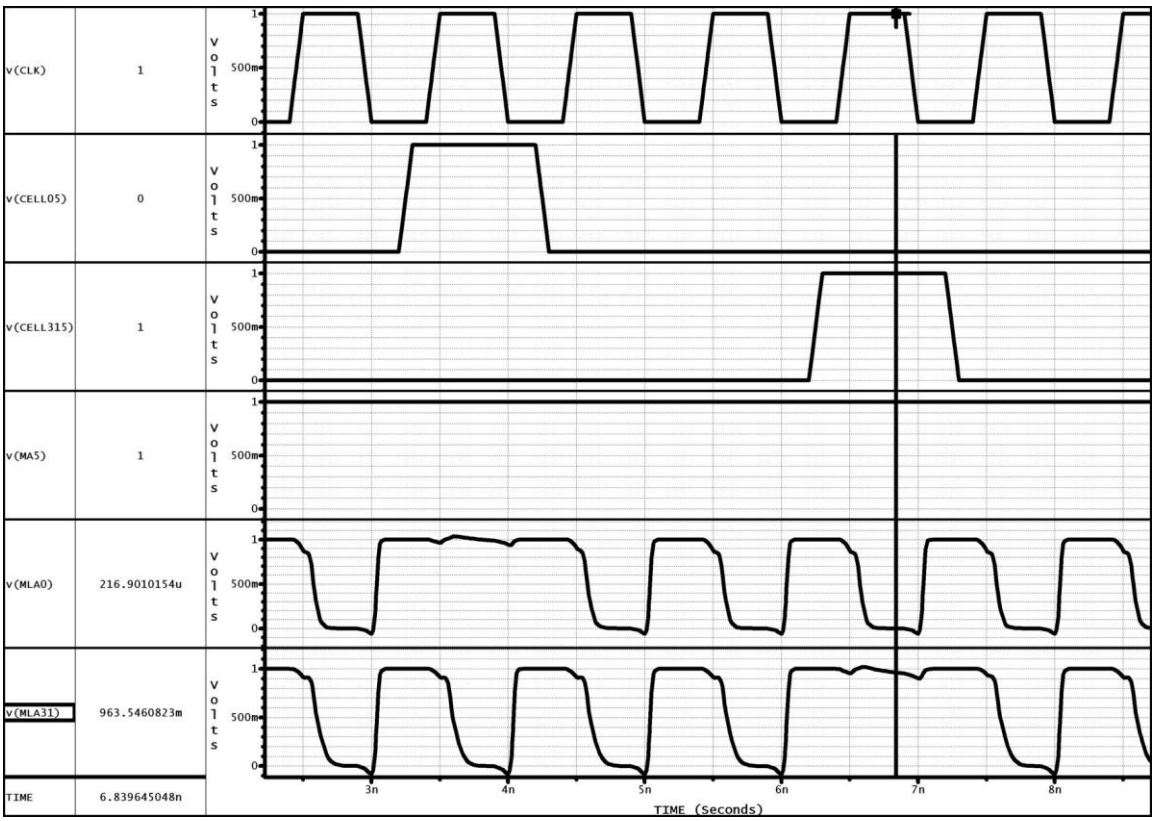


Figure 9: HSPICE simulation for two parallel ports content-addressable operation. Two ports fetch from row “0” and two ports fetch from row “31”, where only the last column is presented, as described.

The simulation of the content-addressable operation in Figure 9 realizes the circuitry of the Alias Table shown in Figure 6, which shows only one port of content-addressable for ease of understanding and brevity of discussion since the other port has a similar operation. The content of the mask register is compared against all rows of the memory array simultaneously, where the row that has the matching contents activates its match line. Subsequently, the active match line permits its associated index address through the Tri-state buffer to propagate to the output port of the content-addressable. The simulation in Figure 9 presents the content-addressable operation, wherein the mask register A is assumed to match the content of row "0" during some time and row "31" during some other time. During the matches, the match line maintains the pre-charge VDD; and thus, enables the Tri-state buffers to release the particular match index value to the output bus of content-addressable. As a result, the output match index bus "CA[0:4]" releases indices "00000" during the match of contents' row "0", and releases indices "11111" during the match of contents' row "31".

The first row of Figure 9 gives the system clock running at 1 GHz, where the second row shows the cells contents Cell(0,5) and Cell(31,5) of only the last cells of rows "0" and row "31" (Column 5) for ease of visuality and brevity of discussion. The third row shows the mask register “A” content of the last bit (MA5). Consequently, the fourth row of simulation shows the match line "0" and "31" of port A (MLA0) and (MLA31), respectively. The MLA0 is high during the match of Cell(0,5) and the high phase of the clock; otherwise, the MLA0 is low during the high phase of clock. Similarly, the MLA31 is high during the match of Cell(31,5) and the high phase of the clock; otherwise, the MLA31 is low during the high phase of clock. Both match lines are high during the low phase of the clock since they are pre-charged during this phase; thus, all match line are high during the low phase of the clock.

The result shows the index is released at less than 0.35 ns, which is very close to the derived content-addressable access time in Eq. (3), that is:

Content-Addressable Access-Time = 0.02 ns/GD * 18 GDs = 0.36 ns.

As a result, the content-addressable operation can safely operate at the half cycle of 1 GHz during the high phase of the clock.

6. RESULTS AND COMPARISONS

The HSPICE simulations conducted in the previous section verifies that the proposed Alias Table exploits the read, write, and content-addressable operations in one cycle without any latency cycle. Such that, the four read ports and two content-addressable ports have occurred in parallel during the high phase of the clock, while the two write ports and pre-charge mechanisms have occurred during the low phase of the clock. Additionally, the simulations showed that the worst operation delay time is less than 20 GDs as a conservative delay measure. Thus, considering a technology factor of 65 nm, the clock cycle of the design can safely run at 1 GHz with a slew rate of 0.1 ns/1V.

Moreover, the Alias Table design has efficient power consumption saving factors. One of the essential factors, the design operates at a 1 V power supply and can further scale for continued CMOS technologies. Additionally, all components have constructed using CMOS transistors with a 65 nm channel length and widths ranging from 3 μm to 5 μm, except for the inverters drivers' widths that are Wp = 10 μm and Wn = 7 μm. Another major contributing factor for low power design is the use of an SRAM-CAM memory array of 21T-Cell structure that is based on the 8T-Cell with a standard geometry size of 65 nm from Intel [25]. The 21T-Cell SRAM-CAM operates at 1V power supply with no sense amplifier obviating a biasing current, which considers a major source of continuous drawn power. Besides, the 21T-Cell provides separate ports for each operation, avoiding a charge contention that minimizes the current density path from power supply to ground, and provides a rail-to-rail noise margin.

TABLE 2 summarizes the comparison of the characteristics between several state-of-the-art Memory Alias Table structures. Key factors such as power consumption, operating speed, performance with latency cycle, operation ports, and cell structure have been discussed and presented. This comparison evaluates the design's factors independent of the underlying technology factor since it is challenging to find comparable designs with the same technical

TABLE 2. Comparison between prior works and the proposed Memory Alias Table design

	Design Structure	**Memory Cell Structure/ ISA**	**Read Ports**	**Write Ports**	**Content-Addressable**	**Characteristics Pros/Cons**	
[7]	Pipeline Design to reduce power and offset low speed	6T-Cell with sensing Amp/ 4-WAY	12-Ports	4-Ports	1-Port Outside memory array	Moderate speed large power Two latencies	Overhead area of pipeline structure to offset low speed
[14]	Design new memory cell, where each cell contains priority	7T-Cell with Pre-charge/ 4-WAY	8-Ports	4-Ports	1-Port Outside memory array	Low Speed Low power One latency	Expensive cell with very large area and slow

	encoders and multiplexers						design
[13]	FIFO memory array equals the number of logical registers.	6T-Cell with FIFO Structure/ 4-WAY	8-Ports	4-Ports	1-Port Outside memory array	High speed Large power Four Latencies	Not efficient in mapping for free list
Our Work	SRAM-CAM based 8T-CELL with separate ports and decoders for each operation	21T-Cell based on 8T-Cell/ 2-WAY	4-Ports	2-Ports	2-Ports Within memory array	High speed Low power No latencies	The design is structured for 2-Way processor and can be expanded for 4-Way processor

Parameters and specifics. However, the comparison still provides insights about relative power consumption, speed, scalability, and design latency throughput. Noticeably, some of the counterpart designs realize content-addressable operation outside the memory array by having a prefix-tree structure of comparators and priority encoders, which worsen the critical path delay and area overhead. In general, designs inherit several latency cycles to compensate for high-frequency operations, and thus, maintain performance from degradation.

The design in [7] is based on the 6T-Cell structure with 4-ports and 12-ports for write and read operations, respectively. Subsequently, the cell has a narrow noise margin between writes and reads, requiring a high-sensitive sense amplifier, which usually draws large biasing currents to leverage the operation speed. Additionally, the design uses an internal pipeline structure in order to remedy the slow clock-cycle for the cost of extra latency cycles. The design further disadvantage harvests large power consumption, and not suitable for continued CMOS scaling technology that requires 1 V or less for the power supply voltage.

The design in [14] operates at the power supply 1 V; however, it is 3X slower than the proposed design. The memory array structure is based on the 7T-Cell that is commonly known for efficient power consumption in the trade of low-speed operation. Each cell contains a comparator and a priority encoder to exploit content-addressable operation; results in a large area cost. The design still requires several latency cycles to offset the low-speed operations; which influence the overall throughput.

The design in [13] exploits a First-in First-out (FIFO) memory array that only reads or writes from the next location. Thus, the design suffers from the low utilization and fragmentation of the memory array. Additionally, the data dependencies of the write and the read operations are accomplished through several priority encoders with multiplexers outside the FIFO array. Furthermore, the content-addressable exploits a prefix-tree structure of comparators outside the FIFO array. The design requires four latency cycles to holds the three operations. Moreover, the FIFO array is based on the 6T-Cell that is known for its low efficiency on power consumption and technology scaling in comparison with the 8T-Cell.

7. CONCLUSION

In this work, the Memory Alias Table circuit design is proposed with size 32-row x 5-bit that exploits two write parallel ports, four read parallel ports, and two content-addressable parallel ports. The high phase of the clock holds the read and the content-addressable operations, while the low phase of the clock holds the write and the pre-charge operations. Therefore, the design operates on a single clock cycle and obviates any latency degradation. The cell of the memory array has an SRAM-CAM structure with 21 transistors. The SRAM portion of the cell carries all advantages of the 8T-Cell SRAM-based and constitutes four parallel read ports and two parallel write ports.

The CAM portion of the cell is based on XOR pass gate logic with a pre-charged match line, which shares with all cells in the same row. Every row has two match ports that receive data for comparison from two mask registers. The mask register broadcasts its content to all rows of the memory array. Concurrently, each row examines its cells' content with the content of the mask register and affect its match line value. If the match line maintains its pre-charge value, then there is a match; and thus, the 5-bit index of that particular row is released. Therefore, content-addressable circuitry precludes a large tree of comparator logic structure. Additionally, the input buffer of the memory array has a priority encoder to alleviate write-after-write data dependencies. In extensive HSPICE simulations, the results show that the clock cycle of 20 standard CMOS gate delays (i.e., independent of technology parameters) can compensate for the three operations without any latency cycles. As a result, the Memory Alias Table operates at clock cycle 1 GHz with a 1 V power supply based on 65 nm technology surpasses most of the current release designs by 3X-to-5X. Furthermore, the proposed design operates with a 1 V power supply and offers continued technology scaling as an attractive feature for low power design.

REFERENCES

[1] K. Patsidis, C. Nicopoulos, G. Ch. Sirakoulis, G. Dimitrakopoulos, “RISC-V2: A scalable RISC-V Vector Processor,” IEEE International Symposium on Circuits and Systems (ISCAS), Sevilla, Spain, Oct. 10-21, 2020.

[2] D. Leibholz and R. Razdan, “The Alpha 21264: A 500 MIPS Out-of-Order Execution Microprocessor,” Proc. Compcon, IEEE Computer Society Press, San Jose, CA. USA, Feb. 23-26, 1997.

[3] K. Moore, Samuel. Breaking the multicore bottleneck. IEEE Spectrum. 53. 16-17. 10.1109/MSPEC.2016.7607015, 2016.

[4] J. L. Hennessy and D. A. Patterson, Computer Architecture: A Quantitative Approach, Morgan Kaufmann; 4th edition, Sept. 27, 2006.

[5] J. P. Shen and M. H. Lipasti, Modern Processor Design: Fundamentals of Superscalar Processors, Waveland Press, Inc., 2013.

[6] M. Postiff;D. Greene;T. Mudge, “The store-load address table and speculative register promotion,” Proceedings 33rd Annual IEEE/ACM International Symposium on Microarchitecture. MICRO-33, Monterey, CA, USA, 2002.

[7] E. Safi;A. Moshovos;A. Veneris, “Two-Stage, Pipelined Register Renaming,”IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 19, Issue 10, pp. 1926-1931, 2011.

[8] D. Sima, “The design space of register renaming techniques,” IEEE Micro, Vol. 20, Issue: 5, pp. 70-83, 2000.

[9] G. Kucuk;O. Ergin;D. Ponomarev;K. Ghose, “Reducing power dissipation of register alias tables in high-performance processors,” IEE Proceedings - Computers and Digital Techniques, Vol. 152, Issue: 6, 2005.

[10] T. N. Buti; R. G. McDonald; Z. Khwaja; A. Ambekar; H. Q. Le; W. E. Burky; B. Williams, “Organization and Implementation of the Register Renaming Mapper for Out-of-Order IBM POWER4 Processors,” IBM Journal of Research and Development, Vol. 49, Issue. 1, pp. 167 – 188, 2005.

[11] S. Petit; R. Ubal; J. Sahuquillo; P. López, "Efficient Register Renaming and Recovery for High-Performance Processors," IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, Vol. 22, Issue 7, pp. 1506 – 1514, 2014.
[12] R. Sangireddy, "Fast and low-power processor front-end with reduced rename logic circuit complexity," IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, May 21-24, 2006.
[13] C. Müller-Schloer, W. Karl, and S. Yehia, "Complexity-Effective rename table design for rapid speculation recovery," In International Conference on Architecture of Computing Systems, Springer, Berlin, Heidelberg, pp.15–24, 2010.
[14] De Gloria, Alessandro, and Mauro Olivieri. "An application specific multi-port RAM cell circuit for register renaming units in high speed microprocessors." Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on. Vol. 4. IEEE, 2001.
[15] A. M S Abdelhadi; G. G. F. Lemieux, "Modular Switched Multiported SRAM-Based Memories," ACM Transactions on Reconfigurable Technology and Systems, Vol. 9, Issue 3, pp. 1–26, July 2016.
[16] Yen-Jen Chang; Kun-Lin; TsaiYu-Cheng; ChengMeng-Rong; Lu, "Low-power ternary content-addressable memory design based on a voltage self-controlled fin field-effect transistor segment," Computers & Electrical Engineering, Elsevier, Vol. 81, pp. 528-540, January 2020.
[17] Saleh Abdel-Hafeez, Shadi M. Harb, William R. Eisenstadt, "Low-Power Content Addressable Memory With Read/Write And Matched Mask Ports", PATMOS 2007, LNCS 4644, Gothenburg, Sweden, pp. 75–85, Sep. 2007.
[18] S. Petit;R. Ubal;J. Sahuquillo;P. López, "A power-aware hybrid RAM-CAM renaming mechanism for fast recovery," IEEE International Conference on Computer Design (ICCD), Nov. 2009.
[19] S. S. Ensan, M. H. Moaiyeri,B. Ebrahimi, S. Hessabi, A. Afzali-Kusha, "A low-leakage and high-writable SRAM cell with back-gate biasing in FinFET technology," Springer, Journal of Computational Electronics (2019), Vol.18, pp. 519–526, March 2019.
[20] A. K. Singh; M. M. Seong; C. M. R. Prabhu, "A data aware 9T static random-access memory cell for low power consumption and improved stability," International Journal of Circuit Theory and Applications, Wiley, Vol.8, Issue 4, Jan. 2013.
[21] G. Prasad; N. Kumari; B. Chandra; M. Ali, "Design and statistical analysis of low power and high speed 10T static random-access memory cell," International Journal of Circuit Theory and Applications, Wiley, Vol. 48, Issue 8, May 2020.
[22] W. Hussain; S. M. Jahinuzzaman, "A 7T SRAM bit-cell for low-power embedded memories," Proceedings of the 21st Edition of the great lakes symposium on Great lakes symposium on VLSI (GLSVLSI '11), pp 121–126, May 2011.
[23] Y. Sharma, A. Singh, A. Pandey, "Comparative Design and Analysis and CMOS SRAM Cell," International Conference on Signal Processing and Communication (ICSC), NOIDA, India, March 7-9, 2019.
[24] H. Zhu; V. Kursun, "A comprehensive comparison of superior triple-threshold-voltage 7-transistor, 8-transistor, and 9-transistor SRAM cells," IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, Australia, June 1-5, 2014.
[25] Saleh Abdel-hafeez and Sarathy P. Sribhashyam, "System and Method for Efficiently Implementing a Double Data Rate Memory Architecture", US patent No. 6356509, March 15, 2002.
[26] A Nand Tech (Intel I7): http://www.anandtech.com/show/2594/10, 2002.
[27] K. Nii, Y. Masuda, M. Yabuuchi, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, M. Igarashi, K. Tomita, N. Tsuboi, H. Makino, K. Ishibashi, H. Shinohara, "A 65 nm Ultra-High-Density Dual-Port SRAM with 0.71um/sup ~/ 8T-Cell for SoC," Symposium on VLSI Circuits, Digest of Technical Papers, Honolulu, HI, USA, 2006.
[28] Y. Chen, M. Fan, P. Hu, P. Su, C. Chuang, "Ultra-low voltage mixed TFET-MOSFET 8T SRAM cell," Proceedings of the 2014 international symposium on Low power electronics and design (ISLPED '14), pp 255–25, August 2014.
[29] M. Patnalaa, A. Yadava, J. Williamsb, A. Gopinatha, B. Nutterb, T. Ytterdalc, M. Rizkalla, "Low power-high speed performance of 8T static RAM cell within GaN TFET, FinFET, and GNRFET technologies – A review," Solid-State Electronics, Elsevier, Vol. 163, January 2020.
[30] Y. Kumar, P. Gupta, "External memory layout vs. schematic," ACM Transactions on Design Automation of Electronic Systems, Vol. 14, Issue 2, pp. 1–20, March 2009.
[31] A. Teman, D. Rossi, P. Andreas, P. Meinerzhagen, L. Benini, A. P. Burg, "Power, Area, and Performance Optimization of Standard Cell Memory Arrays Through Controlled Placement," ACM

Transactions on Design Automation of Electronic Systems, Volume 21, Issue 4, Article No.: 59, pp 1–25, Sept. 2016.
[32] Saleh Abdel-hafeez, Muhanad Quader, sanabel alotoom, "Alias Table Memory Circuit for Register Renaming Unit", IEEE 10th International Conference on Information and Communication Systems (ICICS) , IEEE, Jordan, June 2019.
[33] H. Tabani, Jose-Maria Arnau, J. Tubella, A. Gonzalez, "A Novel Register Renaming Technique for Out-of-Order Processors," IEEE International Symposium on High Performance Computer Architecture (HPCA), Vienna, Austria, Feb. 24-28, 2018.
[34] S. Abdel-Hafeez and A. Matalkah, "CMOS Eight-Transistor Memory Cell for Low-Dynamic-Power High-speed Embedded SRAMS," Journal of Circuits, Systems and Computers, World Scientific, Vol. 17, No. 5, pp. 845-863,Oct. 2008.
[35] S. Abdel-Hafeez, M. Shatnawi, and A. Gordon-Ross, "A Double Data Rate 8t-Cell SRAM Architecture for Systems-On-Chip," IEEE 14Th International Symposium on System-on-Chip, Tampere, Finland, October 11-12, 2012.
[36] H. Jooypa, D. Dideban, "Impact analysis of statistical variability on the accuracy of a propagation delay time compact model in nano-CMOS technology," Springer, Journal of Computational Electronics (2018), Vol. 17, pp. 192-204, Dec. 2017.
[37] Taiwan Semiconductor Manufacturing Corporation, 0.65 μm CMOS ASIC Process Digests, Hsinchu, Taiwan [Online], 2005.
[38] Synopsys, HSPICE, Mountain View, CA [Online], 2016.

AUTHORS

Saleh Abdel-hafeez received his BSEE, MSEE, and Ph.D. in Computer Engineering in the field of VLSI design, the USA. In 1997, he joined S3.inc as a member of their technical staff, where he performed IC circuit design related to cache memory, digital I/O, and ADCs. He has three patents (6,265,509; 6,356,509; 20040211982A1) in the field of IC design. Currently, he is a Professor in the College of Computer and Information Technology, University of Science and Technology, Jordan. His research interests include circuits and architectures for low power and high-performance VLSI. Prof. Abdel-hafeez is a former chairman of the computer engineering department.

Sanabel Otoom received her BSEE and MSEE in Computer Engineering. In 2019, she publishes a paper about Register Alias Table at the 10th International Conference on Information and Communication Systems. Currently, she is seeking an operationality to start her Ph.D. Her research interests include circuits and architectures for high-performance processors and memory.

Quwaider, Muhannad earned his Ph.D. and M.S. at Michigan State University in East Lansing, the USA, and his B.S. at Jordan University of Science and Technology, Jordan. Dr. Quwaider served as vice-dean of the Faculty of Computer and Information Technology from 2018 to 2020 and a chairman of the Computer Engineering department at Jordan University of Science and Technology since 2018. Additionally, He served as a steering committee and TPC chair for the International Conference on Information and Communication Systems (ICICS) from 2012 to 2020. His current research interests include the broad area of wireless data networking, cloud computing, the internet of things, low-power network protocols, high-performance circuit design, and body area network.

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

E-TEACHING AND -LEARNING IN CRISIS SITUATIONS: THEIR EFFECT ON NEW DIRECTIONS OF THINKING IN HIGHER EDUCATION

Nitza Davidovitch and Rivka Wadmany

Ariel University, Israel

ABSTRACT

The COVID-19 year was a difficult and challenging year in all areas of life. The academic world as well was compelled, in a matter of days, to shift from face-to-face learning on campus to e-Learning from a distance, with no adequate preparation. Despite the difficulties generated by e-Learning and students' many complaints, the Israeli Council for Higher Education and institutions of higher education are preparing for a new era, where online courses will constitute an integral part of studies. The purpose of the study was to examine the attitude of lecturers and students to the benefits and shortcomings of e-teaching with its various aspects from a systemic, multi-institutional perspective. The study included 2,015 students and 223 lecturers from different academic institutions: universities, academic colleges of education, academic colleges of engineering, and private colleges.

The research findings show that only one third of the lecturers expressed a preference for e-Learning. With regard to the types of preferred e-lessons: 69% would prefer to teach theoretical classes online, while 42% would prefer to teach exercise classes online. Only 14% would prefer to teach practical classes online, and only 19% would prefer to conduct workshops online. Lecturers were found to have more negative opinions of e-teaching than students: Two thirds of the lecturers (60%) are not happy that e-Learning reduces their interpersonal interaction with the students and among the lecturers. The proportion of lecturers who lament the lack of social interaction is higher than that of students who feel this lack (40%). About two thirds of lecturers noted the lack of social and emotional personal interaction with students and lecturers as one of the main shortcomings of e-Learning. Moreover, most of the lecturers do not perceive e-Learning as an advantage with regard to the quality of teaching and learning and only one third of the lecturers were of the opinion that e-teaching is on a higher standard than face-to-face teaching. Only one sixth of the lecturers were of the opinion that e-Learning is worthwhile for students with regard to their ability to handle the studies and the study material or to gain from the lessons. The study indicates the need for perceptual changes among the lecturers, such that they will reexamine the teaching and learning processes and adjust their role and fields of responsibility to the new opportunities provided by the technological tools and learning environment. The success of e-Learning requires suitable pedagogical educational approaches rather than copying teaching patterns from traditional frontal approaches to online teaching patterns. The research findings indicate the roles of the lecturer in the digital era, and particularly the role of the professionals responsible for teaching and learning in academic institutions, primarily with regard to the pedagogical aspects. The system of academic education has proven that beside the difficulties generated during the crisis, distance learning has many advantages such as the ability to study anytime and anywhere, efficient planning, and adapting the courses and study methods to the students. Nevertheless, the research findings prove that there is no alternative to personal contact, encounters between the teacher and students and among the students. E-learning constitutes a unique and powerful solution, but not an exclusive solution, and it is not necessarily appropriate for all disciplines and teaching and

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 107-114, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111908

learning goals. It appears that combining e-Learning with "face to face" learning can enhance the learning experience, the successes, and students' achievements. Advance preparation, as well as planning a new daily schedule on campus, might advance lecturers and students, in a gradual and structured way, to the challenging tasks of future teaching.

KEYWORDS

E-Learning, academic teaching, COVID-19, crisis situations, higher education

1. INTRODUCTION

We are witnessing significant processes, the transformation of pedagogical paradigms, and new ways of combining technology in processes of pedagogic development and in teaching [1-3]. Researchers note that it is now not up to us, and different institutions and organizations, including institutions of higher education, have no option of disregarding technological processes and the possibilities offered by distance learning. In this paper, we shall review leading and relevant trends affecting many differentglobal spheres (the labor market, the economy, politics, the environment) and we shall address their effect on the world of higher education [4].

There are two basic assumptions for thinking about changes in the world of higher education [5-6]:

1. Higher education does not operate in a vacuum. It is a sphere that is affected by the relevant circumstances. It is necessary to take into account such significant transformations even if they do not occur (at first) in higher education, rather in the surrounding spheres [7].

2. Academia is at present under criticism and strict inspection with regard to its status and contribution to the new world:

   - **Economy of time** – How do students benefit from academic studies? Why is it worthwhile to invest time and money in higher education?

   - **"Face-to-face" encounters are notthe most important**. There are many tools and means that support learning.

   - **Distance learning** – A trend that began decades ago but has accelerated (digital articles and textbooks, audio books, recorded lectures, interactive course websites).
     The option of academic studies without leaving home (including live lessons, asynchronous study material, and online exams) exists at present at all institutions in the world. Institutions are distinguished by the unique mix they generate within studies, between the synchronous and asynchronous components.
     We are required to change the paradigms not only with regard to encounters on Zoom but rather also with regard to the entire learning process, through developing asynchronous components, improving learning environments, ways of evaluation, and others.

   - **Individually adapted teaching.** Differentiation and a response to each student according to his or her level of knowledge and unique learning style. In addition, the introduction of adaptive technologies to studies raises the added value of individual adjustment both when presenting information and in the pedagogical solutions offered: determining the ratio of frontal encounters and Zoom encounters (expanding

time and place), self-study and assignments on the course website, personal assignments, and others.

- Digital pedagogy that defines the goals of education and the image of students in the 21st century in order to adapt them to the dynamic and changing circumstances in which students operate [8].

Soft skills – In an era when knowledge is up to the students and accessible to all, teaching and developing learning skills and other important skills that will allow students to do better at work and in the real world in any field they engage in, is very important [9].

The jobs of the future are still unknown and therefore enhancing skills such as problem solving, creativity, teamwork, time management, networking, and so on, will ensure capabilities relevant for the future world in all areas [10].

Experiential learning – In response to the criticism aimed at academia for its separation from the "field" and the practical world and in the wish to generate meaningful, experiential, and memorable learning for students, there is a demand to integrate more experiential learning. Learning that confronts students with real problems from the field and invites dealing with them, learning that combines familiarization with the practical world of the discipline and forming meaningful experiences. The technological development of virtual reality and augmented reality systems helps form a tangible experience for the students, from home as well [11].

Networked learning – As part of the global perception of transitioning from hierarchies to a network, networked learning is based on the premise that knowledge cannot belong to a single person and that no significant learning is produced by transferring knowledge from a single element to a group. Knowledge must be produced in a group by all its members.In this context, the lecturer not only imparts knowledge but rather mediates it, facilitating group learning [12].

- **Artificial intelligence –** This significant technological trend that has been growing rapidly in recent years, is finally beginning to permeate the world of higher education as well. Systems for managing learning are using the assistance of algorithms that analyze the method of operation utilized by learners in the environment, the difficulties and preferences for various media, and provide learners with adjusted feedback according to this algorithm analysis.

Teaching that encompasses data: Feedback on effectiveness, support activities, the pattern of learning in the course, data on the lesson and on the students – are a basis for determining indicators and recommendations for support activities.

The result: Creating a dynamic of constant improvement.

In the future: Developing an algorithm and analysis of learning patterns to explore learning in the course and individually [13].

The connection between academia and industry – Much criticism has been voiced with regard to the separation between the academic world and the practical and industrial world outside, which students encounter when completing their studies. Institutions of higher education that form a stronger connection between academia and industry will give themselves a relative advantage and a justification for choosing them over the competitors. Such a connection can occur through contents adapted to market needs, links between places of employment, shadowing, boot camps, residency programs for courses and students, and help in placement towards the end of studies [14].

- **Alternative study models**– As part of the premise whereby it is necessary to emphasize the returns and the added value that students receive from academic studies, the understanding that different students have different needs and that academia must generate interesting and diverse alternatives for studies is forming. Not only full degrees, also nanodegrees, training programs, structured programs for certain occupations or professions, and mainly more specific contents that are better tailored to the practical world that follows academia.
- **Students' physical, mental, and social well-being–** This trend raises the need to see the students from a more holistic and humane perspective. Distance teaching reinforces the emotional aspects involved in studies and every self-respecting institution must take into account the consideration of and response to these emotional aspects for the good of the studies. Blurring the boundaries between the home and workplace/place of studies, and the hardships added during the COVID-19 crisis, reinforce the conception of social-emotional learning and the need to grant students the security they need in order to be available for the study process.

- The current era urges people to remain on their toes – **life-long learners**. Learning occurs incessantly, as the changes and dynamic nature of life constantly encourage adjustment to the next thing. Many professions are disappearing and others emerging, requiring us all to leave out comfort zone and to learn at all times according to the needs and to our stage in life. No longer an initial degree and then working 40 years in the same place, rather changing needs, dynamic employment, and diverse studies adapted to the different stages in life.

In summary – what is our added value as an institution and as lecturers?

Institutions of higher education can no longer rest on their laurels and expect to continue existing and flourishing while disregarding their dynamic surroundings. Rethinking study contents, their relevance for the life of the students, new and varied methods of evaluation and learning, harnessing technology to recreate learning processes and adapt them to the learning methods and to learners' emotional aspects, all these will ensure that higher education is a meaningful and relevant element in the current era as well.

2. THE CURRENT STUDY

2.1. Purpose of the Study

To explore the attitude of lecturers to the benefits and shortcomings of e-teaching from different aspects and in a systemic, multi-institutional perspective.

2.2. Research Questions

1) What are the benefits and shortcomings of e-teaching on the cognitive dimension (interest, order and organization, and clarity) as perceived by students and lecturers?
2) What are the benefits and shortcomings of e-teaching on the academic-emotional dimension, with a focus on interpersonal interaction and the availability of the lecturers to the students?
3) What are the main difficulties of students in general, with a focus on deficient resources in particular, with regard to e-Learning?
4) What are the personal preferences of students and lecturers with regard to e-teaching and -learning in general and by: type of lesson, approaches to distance learning and teaching, types of institutions of higher education, study departments, and types of lesson, and what factors affect this perception?

5) Are there differences between students and lecturers with regard to the benefits and shortcomings of e-Learning?

2.3. Research Population

The study included 223 lecturers teaching at various academic institutions: universities, academic colleges of education, academic colleges of engineering, and private colleges. Of these, 51% were men and 49% women. Sixty-two percent held the rank of lecturer rank, 20% the rank of teacher, and 18% the rank of professor. Sixty-five percent were teaching in faculties of social sciences and humanities and 35% in faculties of exact sciences. The study also included 2,015 students from different academic institutions: universities, academic colleges of education, academic colleges of engineering, and private colleges.

2.4. Research Tools and Data Analysis

A questionnaire that included statements related to the impact of e-Learning on the quality of learning and to the benefits and shortcomings of e-teaching and -learning.**The questionnaire** was constructed based on the cognitive-emotional model of best teaching devised by Hativa [15]. According to this theory, a good teacher has a teaching capacity that is comprised of two dimensions:

(1) The cognitive dimension: interest, order and organization, and clarity. Reliability: 0.79.
(2) The academic-emotional dimension: interpersonal interaction and the availability of the lecturers for the students. Reliability: 0.73.

Two other areas explored in the study were based on the model devised by Cohen and Davidovitch (2020):

(3) Worthwhileness and improving the student's learning capacity in e-teaching. Reliability: 0.71.
(4) Personal preference and perception of e-Learning by students and lecturers (by type of lesson, method of study, type of institution, department of studies, student convenience, and resources). Reliability: 0.84.

The factors affecting students' preference for e-teaching as perceived by students and lecturers were also examined.

The questionnaire developed included 43 items that referred to four main areas described. The respondents were asked to rank their answers on a scale of 1 to 5 (where 1 means not at all and 5 means very much). In addition, the questionnaire included questions on personal, family-employment, and academic/teaching background of the students and lecturers in the sample. The findings are presented below:

2.4.1. Student Characteristics: Personal, Family-Employment, and Academic Background

Of all the respondents, 46.8% were men and 53.2% women, where 87.7% were undergraduate students, 39.0% were not employed, 67.7% were single, and 31.1% married. In addition, 89.3% were Jewish, 8.4% with a high socioeconomic status, 68.0% a medium socioeconomic status, and 23.6% a low socioeconomic status. Almost all (90%) noted that they have the necessary resources and tools for e-Learning.

2.4.2. Lecturers' Perception of E-Teaching

Only 31% of lecturers in academia prefer e-teaching. The research findings indicatethat the rate of lecturers who lament the lack of social interaction is even higher than that of students who feel the same way. Most of the lecturers do not perceive e-learning as an advantage with regard to the actual quality of teaching and learning.

2.4.3. Preference for E-Teaching by Type of Lesson

With regard to the preferred type of e-lessons:

- 69% of lecturers prefer to teach theoretical classes online.
- 42% prefer online exercise lessons.
- Only 14% prefer online practical lessons.
- Only 19% prefer to conduct workshops online.

Namely, lecturers have negative views of e-teaching, even more than do students.

Two thirds of the lecturers are not happy that e-Learning reduces their interaction with the students and with their peers. The rate of lecturers who miss the social interaction is even higher than the rate of students who hold this opinion.About two thirds of the lecturers noted the lack of personal, social, and emotional interaction with students and lecturers, as one of the main disadvantages of e-Learning.

Most of the lecturers do not perceive e-Learning as an advantage with regard to the actual quality of teaching and learning.Only about one fifth of the lecturers are of the opinion that e-teaching is on a higher standard than face-to-face teaching.Only one sixth of the lecturers are of the opinion that e-Learning is worthwhile for students with regard to their ability to cope with the learning and the study material or to benefit from the lesson.

3. SUMMARY OF THE FINDINGS

The research findings indicate that:

- **Students and lecturers do not show a high preference for e-Learning during the COVID-19 crisis:** Less than half the students and about one third of the lecturers expressed a preference for studying in this method.
- **Students and lecturers noted the lack of personal interaction with students and lecturers** as one of the main shortcomings of e-Learning/teaching conducted in their institution.
- One of the disadvantages of e-teaching/learning, which arose in the two populations, was the concern that students' score average or the average of lecturer evaluations, would be affected following the transition to e-Learning.
- Saving resources (such as: petrol for traveling to the academic institution, hours of standing in traffic jams, hours of waiting between lessons at the institution) is one of the benefits of e-Learning as perceived by the students.
- Most of the students do not perceive e-Learning as providing them with an advantage with regards to the quality of learning. Notably, improving the ability to study online is the measure that most affects students' preference for e-Learning, but only 39% of students are of the opinion that e-Learning indeed gives them such an advantage.

- Among the population of lecturers, the perceived worthwhileness of e-teaching for the students and the degree of interest, order, organization, and clarity in teaching, are factors that have a considerable impact on the preference for e-teaching, however less than half the lecturers are of the opinion that e-teaching has these qualities. Notably, older lecturers show less preference for e-teaching than do younger lecturers. They may have more difficulty adapting to teaching in this method.
- Both students and lecturers are of the opinion that e-Learning/teaching are not equally suitable for all types of lessons: Both populations are mostly of the opinion that a theoretical lesson can be learned via e-teaching, but few think that it is possible to take a workshop or practical lesson via e-Learning.
- The students prefer e-Learning in the synchronous online lesson approach over asynchronous e-Learning via recorded lectures, and some are of the opinion that the best way of studying online is by a combination of these approaches.
- About one tenth of the students report that there is a problem that prevents them from studying online. The main difficulty noted by the students is the lack of technology that enables implementation of e-Learning (computer, microphone, camera, etc.). Other difficulties indicated by the students are the lack of proper study conditions and the difficulty to adjust to e-Learning. Most of the students (90%) noted that they have the necessary resources and tools for e-Learning.
- The students are of the opinion that use of technology is not done thoughtfully and does not affect the quality of teaching and learning processes. They claim that faculty members use it technically more than pedagogically.

4. DISCUSSION

The study indicates the need for perceptual changes among the lecturers, such that they will reexamine the teaching and learning processes and adapt their role and fields of responsibility to the new opportunities afforded by the technological tools and learning environment. The success of e-Learning requires appropriate pedagogic educational approaches rather than copying teaching patterns from traditional frontal approaches to online teaching patterns. The research findings indicate the roles of the lecturer in the digital era, and particularly the role of the professionals responsible for teaching and learning in academic institutions, particularly in the pedagogical aspects.

The system of academic education has proven that beside the difficulties formed as a result of the crisis, distance learning has its advantages, such as the ability to study at any time and anywhere, efficient planning, adjusting the courses and manners of studying to the students. Nevertheless, the research findings prove that there is no replacement for personal contact and for the teacher-student encounter and encounters between the students. E-Learning is a unique and powerful but not an exclusive solution and it does not suit all disciplines and ways of teaching. The combination of e-Learning with “face-to-face” learning can enhance the learning experience, as well as students’ success and achievements. Advance preparation and planning a new schedule on campusmight boost the progress of lecturers and students in a gradual and structured way, for challenging tasks of teaching in the future.

From theory to practice:

- A need for perceptual change among the lecturers, such that they will reexamine the teaching and learning processes and adapt their role and fields of responsibility to the new opportunities afforded by the technological tools and learning environment.

- The success of e-Learning requires appropriate pedagogic educational approaches rather than copying teaching patterns from traditional frontal approaches to online teaching patterns.
- The system of academic education has proven that beside the difficulties formed as a result of the crisis, distance learning has its advantages, such as the ability to study at any time and anywhere, efficient planning, adjusting the courses and manners of studying to the students.
- The need for personal contact and for teacher-student encounters and encounters between the students. E-Learning is a unique and powerful but not an exclusive solution, and it is not necessarily appropriate for all disciplines and ways of learning and teaching. It appears that the combination of e-Learning with "face-to-face" learning can enhance the learning experience, as well as students' success and achievements. Advance preparation and planning a new schedule on campus might boost the progress of lecturers and students in a gradual and structured way, for challenging tasks of teaching in the future.

REFERENCES

[1] Guri-Rosenblit, Sara. (2010) *Digital technologies in higher education: Sweeping expectations and actual effects*, Nova Science.

[2] Guri-Rosenblit, Sara. (2018) "E-teaching in higher education: An essential prerequisite for e-Learning", J*ournal for New Approaches in Educational Research*, Vol. 7, No. 2, pp93-97. doi: 10.7821/near.2018.7.298

[3] European Commission (2020) *Digital Education Action Plan 2021-2027 – Resetting education and training for the digital age.*

[4] Davidovitch, Nitza. &R. Wadmany (2021) "E-learning in times of crisis – An incidental or facilitative event?" In Z. Sinuany-Stern (Ed.), *Handbook of operations research and management science in higher education* (pp. 453-479). Springer Nature.

[5] Almog, Tamar. & Almog,Oz. (2020) *Academia: All the lies: What went wrong in the university model and what will come in its place*. YediotSfarim. [Hebrew]

[6] Katz, Israel. & Gail. Talshir (Eds.) (2019) *Lighthouse or ivory tower? The Israeli academy: Between challenging openness to defensive seclusion*. Resling. [Hebrew]

[7] World Economic Forum. (2020) *The Future of Jobs Report 2020*. World Economic Forum.

[8] Wadmany, Rivka. (2017) "Digital pedagogues' in the information era", In R. Wadmany (Ed.), *Digital pedagogy – from theory to practice* (pp. 11-16). Mofet Institute. [Hebrew]

[9] Contact North (2020) *A new pedagogy is emerging, and online learning is a key contributing factor* (August 4).

[10] Frankiewicz, Becky.,&Tomas. Chamorro-Premuzic (2020) "Digital transformation is about talent, not technology", Harvard Business Review, Vol. 6.

[11] NeborskyNEgorValentinovich. Boguslavsky Mikhail Victorovich, Ladyzhets Natalya Segreevna&Naumova Tatyana Albertovna(2020) "Digital transformation of higher education: International trends",*Advances in Social Science, Education and Humanities Research*, Vol. 437, pp393-398.

[12] Wahab, Ali. (2020) "Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic," *Higher Education Studies*, Vol. 10(3),16-35.

[13] Nichols, Mark. (2020). *Transforming universities with digital distance education – The future of formal learning*, Routledge.

[14] Pelletier, Kathe.,Malcolm. Brown, D.Christopher Brooks, MarkMcCormack, Jamie Reeves, NicholeArbino, Aras Bozkurt, Steven Crawford, Laura Czerniewicz, Rob Gibson, Katie Linder, Jon Mason, VictoriaMondelli, (2021),*EDUCAUSE Horizon Report Teaching and Learning Edition* (pp. 4-50).

[15] Hativa nira (2015). What does the research say about good teaching and excellent teachers? *Hora'ahBa'academya, 5*, 50-55. [Hebrew]

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

BERT_SE: A PRE-TRAINED LANGUAGE REPRESENTATION MODEL FOR SOFTWARE ENGINEERING

Eliane Maria De Bortoli Fávero and Dalcimar Casanova

Department of Informatics, Technological University of Paraná, Curitiba, Brazil

ABSTRACT

The application of Natural Language Processing (NLP) has achieved a high level of relevance in several areas. In the field of software engineering (SE), NLP applications are based on the classification of similar texts (e.g. software requirements), applied in tasks of estimating software effort, selection of human resources, etc. Classifying software requirements has been a complex task, considering the informality and complexity inherent in the texts produced during the software development process. The pre-trained embedding models are shown as a viable alternative when considering the low volume of textual data labeled in the area of software engineering, as well as the lack of quality of these data. Although there is much research around the application of word embedding in several areas, to date, there is no knowledge of studies that have explored its application in the creation of a specific model for the domain of the SE area. Thus, this article presents the proposal for a contextualized embedding model, called BERT_SE, which allows the recognition of specific and relevant terms in the context of SE. The assessment of BERT_SE was performed using the software requirements classification task, demonstrating that this model has an average improvement rate of 13% concerning the BERT_base model, made available by the authors of BERT. The code and pre-trained models are available at https://github.com/elianedb.

KEYWORDS

Word embedding, Software engineering, Domain-specific Model, Contextualized pre-trained model, BERT.

1. INTRODUCTION

The software development process requires some indispensable activities in its planning phase, such as effort estimates for the project requirements, the selection of adequate human resources for development, the search for reusable resources, among others. Often, the information needed to search for such resources is in text format (e.g. use cases, user stories, bug reports). To obtain this information is a very complex task, considering the intrinsic informality in many software development processes regarding the textual artifacts that are produced in them. This difficulty occurs mainly because, in addition to not having a standard structure, these texts include a diversity of domain-specific information, such as source code, links, IP addresses, among others.

A very common aspect in these texts is the occurrence of different words, but they are used in the same context, in which case there is the need to consider them similar because, although the texts are different, their context is similar. In this case, a context can be defined as the text that precedes and/or follows a particular word, sentence, or text, and that contributes to its interpretation and meaning [1]. The context is directly related to the semantics of a word concerning the situation in which it is applied. Therefore, the need to recognize domain specific

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 115-130, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111909

terms and their meaning in each situation in which are applied is highlighted in SE, aiming to obtain the most accurate feedback possible in carrying out various tasks of area.

In this paper, the specific-domain terms are a set of common words in a particular area (e.g. medicine, software engineering), among them, there are strong semantic relations. Some studies have already been made to obtain similar words in SE area [2], similar to WordNet, which consists basically in a thesaurus, grouping words based on their meanings [3], which remain static until they are updated. An example of the need for context representation in SE would be for the distinction of ambiguous words, like in the sentence: ”The implementation depends on the language”. In this case, it is necessary to consider the target word (language) as its context, to infer that this is a programming language and not a speaking language.

Considering the nature of texts in SE, the application of text representation methods based on the characteristics of each word (e.g. bag-of-words) has several limitations. These limitations interfere in the generation of a learning model and are caused by: sparsity, high dimensionality, and as a result, overfitting. Furthermore, this characteristics type is not sufficient to discriminate against each requirement. It occurs because the software textual requirements require a deep analysis, in which, besides the individual characteristics of each word, semantic characteristics are obtained. It means to analyze the meaning of words contained in a requirement related to your context.

Actually, the word embedding models are a strong tendency in NLP. The first word embedding model that utilized artificial neural networks was published in 2013 [4] by Google researchers, and since then, this concept has been part of majority of research in NLP. The word embedding methods aim to mainly capture the semantic of a particular word in a specific context. This method allows the words to be represented in a dense way and with low dimensionality, facilitating applications of machine learning in which NLP is applied. The innovations in the train of word embedding models for a variety of purposes have begun in recent years with the emergence of Word2Vec [4] and GloVe [5], allowing models to learn from rather bulky corpus. So, the contextual representations by means of embedding models have been very useful in the identification of context-sensitive similarity [6], disambiguation of the word meaning [7], [8], induction of the word meaning [9], lexical substitution for the creation of an embedding generic model [10], sentence complementation [11], among others.

The methods to obtain embedding have been considered one of the greatest advanced, and prominent in the area of NLP in recent years. Different methods for the generation of embedding from texts have been developed [12], which is possible to classify them into [13]: context-less or static, and contextualized or dynamic. Contextual embedding has been considered a revolution in NLP. This approach produces different vector representations to the same word in a text, which varies according to its context and, therefore, they are able of capturing contextual semantics of ambiguous words [14], in addition to addressing polysemy issues. In this way, each occurrence of a word is mapped to a dense vector, considering specifically the surrounding context.

Some studies using embedding have been carried out specifically in the field of SE, such as the recommendation of specific-domain topics in Stack Overflow question tags [15], the recommendation of similar bugs [16], sentimental analysis in software engineering [17], ambiguity detection in requirements engineering [18], among others. None of them used contextualized embedding. Therefore, within this new paradigm, the Bidirectional Encoder Representations from Transformers (BERT) [19] have been one of the algorithms which presented the best results in NLP tasks, besides being open-source. With all these benefits, this model has been widely used in various NLP applications. Its use has occurred via pre-trained

models and is available by the authors [19]). These models were pre-trained in a large corpus of unlabeled texts and of general-purpose (e.g. Wikipedia).

The pre-training of language models proved to be highly effective in learning language universal representations from unlabeled data on a large scale [20]. Therefore, there is no need of train from scratch, due to fine-tuning techniques [21]. According to the authors, this technique consists in tuning a generic pre-trained model, using an unlabeled data corpus in the domain of a particular application. This process allows to economize many hours of training and spare the need of the specific rather bulky corpus. Although there are many researches around the application of word embedding in several areas, until where it is known, so far there is no effective contextualized pre-trained model for the SE area. Therefore, this study suggests the generation of a contextualized pre-trained embedding model, able to recognize specific, and relevant terms of the area. Thus, it becomes possible to identify the specific semantic of each parsed sentence. In addition, the existence of this model seeks to help in pattern recognition among these texts, allowing its effective application in several machine learning tasks in the area, which are based on textual data (e.g.bug classification, software effort estimation based on analogy, and others).

Thus, first, the fine-tuning of the generic BERT embedding model, made available by its authors, was performed. Next, this same pre-trained model went through the fine-tuning process, used for that, a specific corpus of SE domain. The result of the fine-tuning process was a contextualized pre-trained model for SE (BERT_SE). Hence, comparative tests were performed between both models: generic and adjusted. The implementation of the approach was divided into three main steps: (1) collection and pre-processing of the used corpus; (2) preparation of data for pre-training, and (3) application of the pre-training method. Then, the evaluation of the results obtained was carried out focusing on identifying if the similarities among sentences of the context of SE, are better expressed by a generic BERT embedding model or a BERT embedding model adjusted (BERT_SE). The preliminary results obtained are promising, motivating the continuity of the research on this topic.

The following content of this paper is organized as follows. Section 2 presents the background containing relevant aspects related to word embedding as well as related works that use embedding models in specific domains. Section 3 presents the construction approach, describing the necessary steps for its implementation and the evaluation metrics used. Section 4 presents the experimental results obtained, followed by the discussion of statistics and model performance. Section 5 presents the threats to validity, followed by the conclusion and future work (session 6).

2. BACKGROUND

2.1. Pre-trained Embedding Models

The pre-training of the language model proved to be highly effective in learning universal representations of language from unlabeled data on a large scale [20]. Among the main benefits of pre-trained language models, is the fact that there is no need to train them from scratch. This characteristic, besides reducing considerably the need for computational cost, saving a lot of hours of training, become unnecessary a highly representative corpus. This is possible through fine-tuning techniques. The fine-tuning approach, also named transfer learning in some contexts, consists in introduce minor parameters of a specific task and train it in the following tasks, simply adjusting all the pre-trained parameters [19] in a generic rather bulky corpus.

Thus, pre-training has been widely applied in various NLP tasks, bringing many benefits and great advances, especially in tasks that have limited data for training [22], [21]. There are two main methods applied in the generation of pre-trained embedding: context-less and contextualized.

Word2Vec, as well as other similar models (e.g. Glove [5]) are considered algorithms for textual representations context-less. This means that these models present restrictions regarding the representation of the context of words in a text, impairing tasks at the sentence level or even fine-tuning at the word level. This is because these models are unidirectional, that is, they consider the context of a word only from left to right, with no mechanism that detects if a particular word has already occurred in the corpus before. Therefore, these models provide a single representation, using a dense vector, for each word in a text or set of texts.

In addition, according to its authors [4] these models are considered very shallow, as they represent each word in only one layer, and there is a limit on the amount of information they can capture. Finally, these models do not consider the polysemy of words, that is, the same word being used in different contexts can have different meanings (e.g. bank – monetary sense; bank - to sit), which is not treated by these models. Another characteristic that is not treated is the ambiguity of the words, that is, when two or more different words have the same meaning (e.g. create, implement, generate).

On the other hand, many advances have occurred in the area of NLP in recent years. Such advances are due, mainly, to deep learning techniques [23]. Among these advances is the possibility of obtaining contextualized embedding. This approach produces different vector representations for the same word in a text, which varies according to its context. Therefore, these techniques are capable of capturing contextual semantics of ambiguous words [14], as well as addressing polysemy issues. From this new paradigm, recent studies have turned to research that applies contextualized embedding models [24] [14], leaving aside the original paradigm, in which there was only one vector of embedding for each single word in one text/set of texts. Thus, each occurrence of a word is mapped to a dense vector, specifically considering the surrounding context.

This representation approach is easily applicable to many NLP tasks, where the inputs are usually sentences and therefore, the context information is available, such as textual software requirements. This new language representation paradigm originated from several ideas and initiatives that emerged in NLP in recent years, such as: coVe [25], ELMo [24], ULMFiT [21], CVT [26], Context2Vec [10], BERT [19] and Transformer OpenAI (GPT e GPT-2) [27]. The BERT contextualized pre-trained model [19], has presented results greatly improved in NLP tasks, and has therefore been widely used in several applications. Its application has occurred through pre-trained models and available by its authors (e.g. BERT_base e BERT large [19]).

2.2. BERT

The BERT is an innovative method, considered the state of the art in pre-trained language representation [19]. BERT models are considered contextualized or dynamic models, and have shown much-improved results in several NLP tasks [22], [24], [27], [21] as sentiment classification, calculation of semantic tasks of textual similarity and recognition of tasks of textual linking.

This model originated from various ideas and initiatives aimed at textual representation that have emerged in the area of NLP in recent years, such as: coVe [25], ELMo [24], ULMFiT [21], CVT [26], context2Vec [28], the OpenAI transformer (GPT and GPT-2) [27] and the Transformer

[29]. BERT is characterized as a dynamic method, mainly because it has an attention mechanism, also called Transformer [19], which allows analyzing the context of each word in a text individually, including checking if each word has been previously used in a text with the same context. This allows the method to learn contextual relationships between words (or subwords) in a text. BERT consists of several Transformer models [29] whose parameters are pre-trained on an unlabeled corpus like Wikipedia and BooksCorpus [30]. It can say that for a given input sentence, BERT "looks left and right several times" and outputs a dense vector representation for each word. For this reason, BERT is classified as a profoundly two-way model because it learns two representations of each word, one on the right and one on the left, and this learning to repeat n times. These representations are concatenated to obtain a final representation to use in future tasks.

The pre-processing model adopted by BERT accomplishes two main tasks: masked language modeling (MLM) and next sentence prediction (NSP). In the MLM task, the authors argue [19] that it is possible to predict a particular masked word from the context. For example, let's say we have a phrase: "I love reading data science articles." We want to train a contextualized language model. In this case, you need to replace "data" with "[MASK]". It is a token to indicate that it is missing. We will then train the model so that it can predict "date" as the missing token: "I love reading articles from [MASK] science".

This technique aims to make the model learn the relationship between words, improving the level of learning, avoiding a possible "vicious cycle", in which the prediction of a word to base on the word itself. Devlin et al. [19] used 15-20% of words as masked words.

The task of NSP is to learn the relationship between sentences. As with MLM, given two sentences (A and B), we want to know if B is the next sentence after A in the corpus or if it would be any sentence.

With this, BERT combines the pre-training tasks of both tasks (MLM and NSP), making it a task-independent model. For this, their authors provided pre-trained models in a generic corpus but allowing fine-tuning. It means that instead of taking days to pre-workout, it only takes a few hours. According to the authors of BERT [19], a new state of the art has been achieved in all NLP tasks they have attempted (e.g. Question Answering (QA) and Natural Language Inference (NLI)).

2.3. Work-related to the use of specific domains

In recent years, several initiatives making use of textual representations have been applied in several domains. But some areas require representations of words that consider particularities of a particular domain (e.g. health, technology, software engineering). The following will present some of the studies that involve the representation of textual data from specific domains.

A study by [31] addresses the task of extracting events from a representation model of a domain-specific dataset. Is described a set of participants (i.e. attributes or roles) whose values are text excerpts. The authors show that learning word representations from unlabeled domain-specific data and using them to represent event roles enable them to outperform previous state-of-the-art event, extraction models.

Another application occurred in biomedical text mining. In this area, there are many entities and syntactic parts that present rich domain information. In this way, [32], presented a model of word embedding specific to this domain.

Focusing on the ES area will be presented in the sequence some research that has explored aspects of specific domains. The approach of a recommendation system of similar libraries of software implementation (e.g. similar to JUnit) was proposed by [15]. This approach solves queries about these libraries by combining the word embedding technique and domain-specific knowledge extracted from millions of Stack Overflow tags. Still in this line, [2] proposed SEWordSim. It is a lexicon, that is, a dataset of similar words specific to the SE domain. The similarity characteristics between the words were extracted automatically from the questions and answers available in Stack Overflow.

Another study proposed by Celefato et al. [17] addresses the problem of applying feelings analysis to the discipline of software engineering. This classifier uses a set of semantic resources based on a domain-dependent lexicon. Ye et al. [33] explored word embedding to improve information retrieval in software engineering. Its ultimate goal is to eliminate the lexical gap between code fragments and natural language descriptions that can be found in tutorials, API documentation, and bug reports. They empirically demonstrate how exploring word embedding improves next-generation approaches to bug tracking.

Word embedding techniques were also applied to estimate the degree of ambiguity of words typical of the context of computer science (e.g. system, database, interface) when used in different application domains [18]. The results show that it is possible to identify variations of the meaning of the terms of computer science in the applied domains, providing an estimate of the distance between the considered domains. A new approach to recommending similar bugs was proposed [16]. The approach combines standard information retrieval techniques and word embedding techniques. Sugathadasa et al. [34] proposed new measures of semantic similarity aimed at specific domains. This measure was created by the synergetic union of word2vec and lexical-based semantic similarity methods.

There is a wide variety of studies that use embedding models, but our proposal differs from these approaches because, firstly, it aims to overcome the limitations imposed by bag-of-words models, especially concerning dimensionality and sparsity. Also, the differential of the proposed study is in its application. Because it offers the pre-training word embedding model and allows a multitasking representation of textual artifacts to perform tasks involving NLP in the SE domain. According to Chen et al. [15], multi-task learning requires that tasks be trained from scratch at a time, which makes it inefficient and often requires careful consideration of the task's specific objective functions. Thus, one of the great advantages of this model is that there is no need to train a model from scratch, or even have a massive corpus representing words and their semantic relationships within that domain.

3. PRE-TRAINED MODEL FINE-TUNING PROCESS

The main objective of this article is to present BERT_SE, a pre-trained language representation model and adjusted for the SE domain. To do this, we first started BERT_SE with the standard weights of BERT_base [19], which was pre-trained in a general domain corpus (Wikipedia in English and Books Corpus).

As a generic model of BERT, the *BERT_base uncased* was applied, which was previously trained and became available by its authors [19] for free use in NLP tasks. Table 1 presents the specifications for this model.

Table 1. BERT pré-trained model used in the proposed approach.

Pre-trained model	Specification
BERT base	BERT_base uncased: 12 layers for each token, 768 hidden layers, 12 heads of attention, 110 million parameters. The uncased specification means that the text was converted to lower case before tokenization based on WordPiece, in addition, removes any accent marks. This model was trained with English texts (Wikipedia) with lowercase letters.

The BERT_base model, as well as the other pre-trained BERT models, offers 3 components [19]:

- A TensorFlow checkpoint (*bert_model.ckpt*) that contains pre-trained weights (consisting of 3 files).
- A vocabulary file (*vocab.txt*) for mapping WordPiece [35] for word identification.
- A configuration file (*bert_config.json*) that specifies the model's hyperparameters.

3.1. Datasets Used for the Fine-tuning

The BERT_base fine-tuning process is performed using a corpus of the SE domain, here called corp_SE. The composition of the corp_SE is shown in Table 2.

A Stack Overflow dataset was chosen because it is relatively restricted, contains a large number of posts and associated tags, and the data is easy to obtain. It is restricted because it refers to a specific domain (of software engineering), that the authors (users of the forum) can use, which is different from a less restricted domain (e.g. Twitter). The remaining corpus used is all derived from open-source software projects or from software development companies that have authorized their application in research in the area. These data were in a .csv file. After that, a basic pre-processing was carried out, with the objective of excluding special characters, HTML tags, and numbers. Thus, in order to maintain a standard, the same pre-processing applied to software requirements data, obtained from open source projects and used by Choetkiertikul et al. (2018), was performed. For the pre-processing, a method based on specific regular expressions was applied. Table 3 presents some examples of sentences that composed the corp_SE.

Therefore, the corp_SE is composed of 456.500 texts, in this paper called sentences. Each sentence has an average length of 61 words. The vocabulary generated by the corp_SE is composed of 1.179.501 words.

3.2. Performing of Fine-tuning

It stands out that the fine-tuning process consists of the use of a pre-trained embedding model from a generic dataset in an unsupervised way, which is adjusted, that is, retrained on a known dataset that is specific to the area of interest. In this case, the fine-tuning was performed on the generic model BERT_base, using corp_SE (according to Table 2). The fine-tuning process of the pre-trained BERT model consists of two main steps [19]:

1) *Preparation of data for pre-training:* initially, the input data is generated for pre-training. This is made by converting the input sentences into the format expected by the BERT model (using *create_pretraining_data* algorithm). As BERT can receive one or two sentences as input, the model expects an input format in which special tokens mark the beginning and end of each sentence, as shown in Table 4. In addition, the tokenization process needs to be performed. BERT provides its own tokenizer, which generates output as shown in Table 5.

Table 2. Dataset used in the composing of corp_SE.

Data source	Specification
Sentence subset of Stack Overflow (www.stackoverflow.com)	A subset of Stackoverflow sentences taken from the Kaggle [36] repository, totaling 137,474 sentences after pre-processing. These sentences consist of posts made by users about doubts and problems related to the most diverse software development technologies.
Software requirements (user stories) obtained from open-source projects	Textual corpus of 319.026 requirements from 16 large open-source projects in 9 repositories (Apache, Appcelerator, DuraSpace, Atlassian, Moodle, Lsstcorp, Mulesoft, Spring, and Talendforge) [37] and from others 22 open-source datasets [38]. According to the authors that available the datasets, all were obtained online or from software companies with permission for dissemination.

Table 3. Example Sentences that Composed the corp_SE (before pre-processing).

Text ID	Text
1	Create project references property pagehtml create property page for project which allows manipulation of embedded references. user can add or remove references from the filesystem or url (if possible).
2	android: permissions failure in android.calendar drillbit testlooks like the android.calendar test recently started failing due to missing permissions, log:code permission denial: opening provider com.android.providers.calendar.calendarprovider2 from processrecord(423b2048 20514:org.appcelerator.titanium.testharness 10082) (pid=20514, uid=10082) requires android.permission.read calendar or android.permission.write calendar code
3	ws security signature support for ws consumer: we should support ws security signature and verification capabilities in ws consumer.

2) *Application of the pre-training method:* the method used for pre-training by BERT (run pretraining) became available by its authors. The necessary hyperparameters were informed, the most important being:

- *max_seq_length:* defining the maximum size of the input texts (set at 100).
- *batch-size:* maximum lot size (set at 32, per use guidance of the pre-trained model BERT base.
- *epochs number:* the standard epochs number of model is 100. This number has even been varied to 500 and 1000 during the experiments.

Table 4. Example of formatting input texts for pre-training with BERT.

Entry of two sentences	Entry of a sentence
[CLS] The man went to the store. [SEP] He bought a gallon of milk.[SEP]	[CLS] The man went to the store.[SEP]

Table 5. Example application of tokenizer provided by BERT.

Input sentence	”Here is the sentence I want embedding for.”
Text after *tokenizer*	[’[CLS]’, ’here’, ’is’, ’the’, ’sentence’, ’i’, ’want’, ’em’, ’##bed’, ’##ding’, ’##s’, ’for’, ’.’, ’[SEP]’]

Justified that in transfer learning is not suggested change in the values of the hyperparameters, except for the *epochs number* of training. The *max_seq_length* was not varied, as it was observed

from the exploratory analysis of the data, that a value equal to 100 would correspond to most sentences in the dataset. For *batch-size* we chose to leave the default value of 32, considering the amount of memory available. The study by Devlin et al. (2019) however, demonstrates that changes in this hyperparameter usually do not lead to large differences in performance. For the generic BERT model, we opted for its version *Uncased L-12* base, here called BERT_base (Table 1). The fine-tuning process for the BERT model was performed as shown the Figure 1.

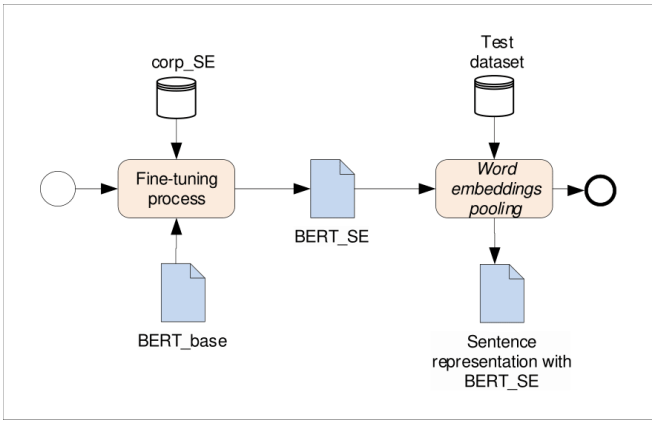


Figure 1. Pipeline of the fine-tuning process of the BERT_base model and generation of the textual representation for the corp_SE.

The entire process, from data preparation to fine-tuning the BERT model, used the algorithms available in the repository *https://github.com/google-research/bert*, in which the authors [19] provides the full framework developed in the Python language.

After performing the fine-tuning, a new pre-trained model is available, as shown in Table 6, which will compose the experiments. It is noteworthy that the proposed model requires pre-training only for the embedding layer. This allows, for example, this pre-trained model is available for other software engineering tasks, or even for different effort estimation tasks. Thus, this pre-trained model may undergo successive adjustments, according to the need of the task to which it will be applied.

3.3. Performing of Fine-tuning

This step consists of analyzing the similarity between a source sentence and a target sentence. The use of the cosine similarity measure was defined considering that it is a predominant way of estimating the similarity of two documents based on word incorporation. Thus, the cosine similarity measure must be applied to the two centroids obtained from the embedding vectors associated with the words in each document [39].

That is, given a sentence set regarding SE, was observed the cosine distance of each sentence concerning the others.

The similarity among sentences is given by the cosine distance among mean embedding vectors that represent them. This is made as to the generic model (BERT_base), as the adjusted model (BERT_SE) obtained from the trained model. That is:

$$t_i = \frac{1}{n}\sum_{i=1}^{n} p_i$$

where p = {p_1, p_2, ..., p_n} is a word of the sentence t = {t_1, t_2, …, t_n} with *n* elements in a vector. So, the cosine similarity of two sentences is given by:

$$cos(\mathbf{t}, \mathbf{t}') = \frac{\mathbf{t} \times \mathbf{t}'}{||\mathbf{t}|| \times ||\mathbf{t}'||}$$

Where both *t* and *t'* are the mean embedding of sentences.

The cosine distance returns a value between [0; 1]. Values closer to 1 indicate greater similarity among vectors. Specialists in the field evaluated, case by case, whether the sentence of origin had semantic similarity or not. For this, a sample of 30 professionals in the field, working in different companies in the field, was selected (e.g. analysts, developers, project managers).

4. RESULTS AND DISCUSSIONS

The results presented in this section illustrate that a general-purpose embedding model, even though it presents a larger vocabulary, does not necessarily adequately represent area specific terms, such as SE. The following experiments intended to demonstrate that results of textual classification tasks in the SE domain could be improved if the language model used for its representation is fined using a specific context dataset.

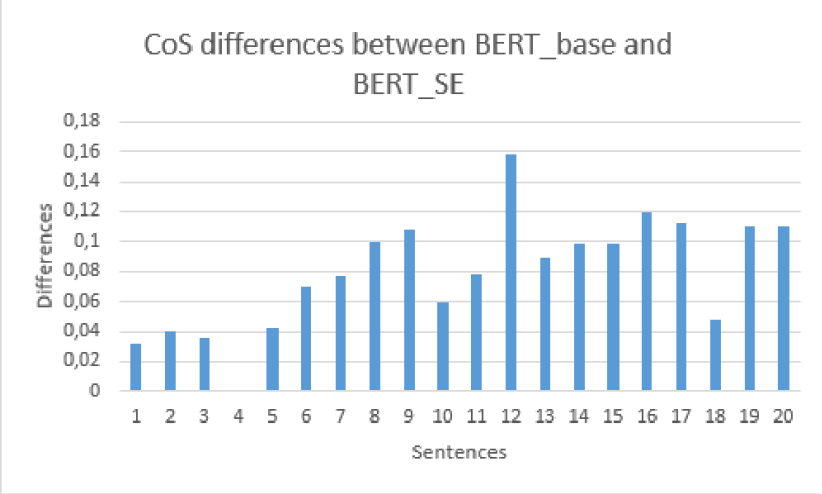


Figure 2. CoS differences between BERT_base and BERT_SE, when the sample S5 (Table 7) is compared with the remaining test sentences.

It's important to remember, that the cosine measure is a trigonometric function that provides a value equal to 1 if the understood angle is zero, that is, if both vectors point to the same place (identical objects). For any angle other than 0, the cosine value is less than one. If the vectors were orthogonal, the coSine would cancel out, and if they pointed in the opposite direction, its value would be -1. Thus, the value of this metric is between -1 and 1. Therefore, the closer the coS value gets to 1, the greater is the similarity between two sentences.

The coS values for BERT_SE are larger, and therefore, closer to 1 when compared to the same values for BERT_base. This indicates a greater cosine similarity between the sentences represented by BERT_SE. This fact can be verified in Table 7, where a BERT_SE improvement rate concerning a base of BERT can be verified.

Especially in software engineering, it is important to note the similarity of contexts among sentences. For example, in S4 (Table 6), both sentences deal with adding a new record (sales order and users). In this way, regardless of the existing content in a register, they are close to implementation operations. Similarly in S3, where both sentences refer to adding a button to a form. Therefore, the coS is expected to be high, since the source and target are from the same context. When presenting the results obtained in Table 6, and asked about the similarity between the sentences used in the evaluation of the model, the specialists approached indicated that there was similarity.

Table 6. Cosine similarity (coS) between sentences in the SE domain, obtained from BERT_base e BERT_SE.

Sample ID	**Source sentence**	**Target sentence**	**coS BERT_base**	**coS BERT_SE**	**BERT_SE improvement rate (%)**
S1	Create user registration allowing to Include a photo and digital	Create a button that allows you to retrieve the last record deleted from the order	0.59	0.71	20.33
S2	Create a method that allows the user to customize sales reports	List all store products by category	0.71	0.84	15.4
S3	Include calculation button by product in the sales order	Create a button that allows you to retrieve the last record deleted from the order	0.65	0.80	18.75
S4	As a salesperson, I want to include sales orders	Create user registration allowing to include a photo and digital	0.58	0.70	20.7
S5	Add user authentication function for accessing the system	As an administrator, I need to have access to a sales report to find out how much I received in a given period	0.75	0.84	10.12

Figure 2 shows an experiment where the difference of coS similarities of both BERT_SE and BERT_base models is measured. If the difference is a positive value then BERT_SE has a high similarity value. If the difference has a negative value then BERT_base shows a better similarity representation. In this experiment, we measure the similarity of sentence S5 (see Table 6) against all other sentences. It is observed that all similarities increases when using BERT_SE model, except for the fourth sentence, where BERT_base and BERT_SE produced equal values.

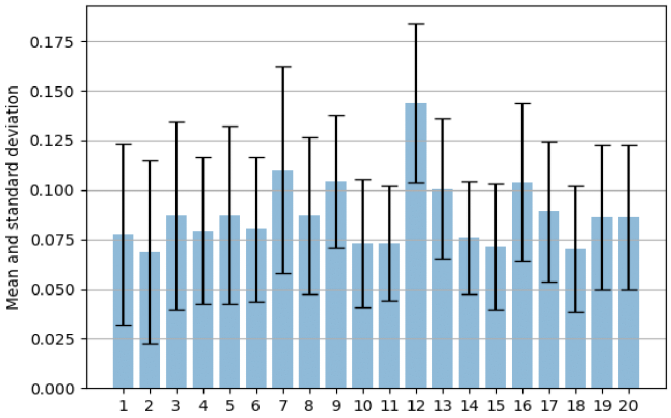


Figure 3. Mean and standard deviation for CoS differences between BERT_base and BERT_SE, for all twenty test sentences.

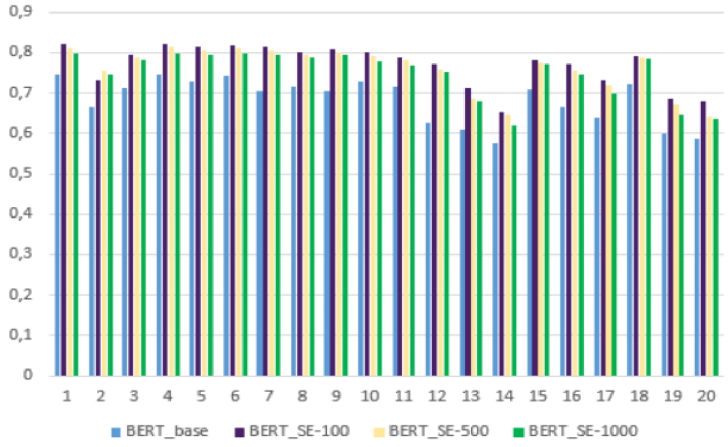


Figure 4. CoS differences mean obtained from BERT_base and BERT_SE (adjusted with different epochs number), for all twenty sentences.

An experiment similar to the previous one is shown in Figure 3. In this figure, the mean and standard deviation of the differences between the representation given by BERT_base and by BERT_SE (trained with 100 epochs), for each sentence is shown. It is observed that the average values of the differences are positive in all sentences, favoring the representation given by BERT_SE. The standard deviation, when compared to this gain, is low, which confirms this positive difference.

This result confirms that the BERT_SE increases the coS similarity if compared with BERT_base. In the next experiment, we investigate if more training time can reach even better results.

Figure 4 shows the average of the cosine distances for each sentence. First when using the BERT_base model and then when applying the adjusted model BERT_SE, trained with a different epochs number (100, 500, 1000). It is observed that the increase in the epochs number of training did not generate significant improvements in the results, but it was evident that the results are always better when applying BERT_SE.

Figure 5 shows the percentage of improvement obtained about the representation of sentences by the BERT_base model when compared with the cosine similarity values obtained when applying the BERT_SE model. It is observed that there was an improvement in all representations of sentences that used the BERT_SE model. The average improvement rate is 13% compared to the initial representation by BERT_base.

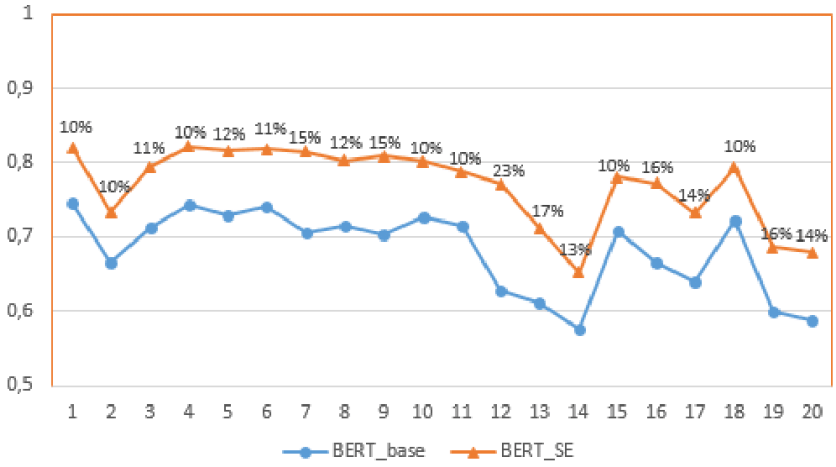


Figure 5. Rate of improvement obtained when representing each sentence with the BERT_SE model in relation to the representation as BERT_base.

5. THREATS TO VALIDITY

This paper proposes the BERT SE, a contextualized pre-trained model to the textual representation in the SE domain. The results of the proposed model were positive, by testing the efficiency in the identification of similar sentences in SE. The pre-trained embedding model BERT_base, which became available by its authors [19], was generated from a dataset extremely wide, containing texts from diverse areas (e.g.Wikipedia), which makes it rather generic. Therefore, it is important to consider the overall model, which is related directly to the availability and diversity of domain data.

To generate a pre-trained embedding model and adjusted it for a specific domain, such as BERT_SE, it is necessary to have a dataset containing only texts related to the SE area, which must be in the same language and pass through the same pre-processing. For the case of SE, this data must reflect the reality of software projects in different areas, different models of the development process, forms of representation of user requirements, technologies, among other attributes. Therefore, we believe that the samples may not be sufficient to represent all the textual variations that exist in SE.

So it is recommended to update BERT_SE whenever new textual data is obtained that is appropriate. Therefore, SE is an area in constant evolution, in which new technologies often appear, another reason to keep the model updated periodically.

The tests performed with the BERT_SE model were validated by a specialists sample in the field, made up of developers, analysts, and project managers. Thus, it is necessary to consider the subjectivity intrinsic to this evaluation, which was carried out according to the opinion and experience of each one of them, which can generate a bias in the results obtained.

6. CONCLUSIONS

The use of pre-trained models for specific domains has shown good results in their applications, such as SciBERT [40] - a pre-trained Language Model for Scientific Text, and BioBERT [41] - a pre-trained biomedical language representation model for biomedical text mining. Therefore, this article aims to explore a BERT_SE, a contextualized pre-trained language model based on BERT, which is destined for textual classification in the field of software engineering.

The BERT_SE was generated from fine-tuning of the BERT_base model [19], a pre-trained embedding model from the generic dataset in an unsupervised way. Thus, BERT_SE was retrained in an unlabeled and specific dataset for the SE area. In this case, fine-tuning was performed on the generic BERT_base model, using the corpus corp_SE. This fine-tuning process to generate BERT_SE, confirmed the statement by the authors of BERT [19] that fine-tuning takes only a few hours on a GPU and does not require a very large specific corpus.

Thus, when compared to the pre-training process, fine-tuning is relatively inexpensive. The results are verified in a sentence representation experiment over software requirements. In this experiment, we expect that a sentence needs be more similar to each other if it belongs to the same context. The results show that the BERT_SE model surpasses the generic model in all representation tests performed, which results in an average improvement rate of 13% about initial representation, given by BERT_base.

As the results of the article show, BERT_SE behaves very well in classifying sentences in the SE area, even considering their context. Thus, it would be possible to perform several NLP tasks in this area with greater precision (e.g. bug classification, software effort estimation based on analogy, and others). The work by Fávero et al. (2020) [42] – in review by the international journal of the area - presents an application scenario for the software effort estimation analogy-based using BERT_SE, where the fine-tuning was performed with a specific less bulky dataset, and the results were positive.

As future work, we intend to launch a version of BERT_SE similar to BERT_large [19]. Besides, increase the volume of the corp_SE and generate a model with its vocabulary, to better correspond to the training corpus, and compare it with the original BERT model. It is also intended to evaluate the model in other NLP tasks (e.g. Named Entity Recognition (NER), Question Answering (QA), etc.) and that may apply to software engineering.

REFERENCES

[1] M. A. K. Halliday, C. Matthiessen, and M. Halliday, An introduction to functional grammar. Routledge, 2014.
[2] Y. Tian, D. Lo, and J. Lawall, “Sewordsim: Software-specific word similarity database,” in Companion Proceedings of the 36th International Conference on Software Engineering. ACM, 2014, pp. 568–571.
[3] C. Fellbaum, “Wordnet,” The Encyclopedia of Applied Linguistics, 2012.
[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[5] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[6] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, “Improving word representations via global context and multiple word prototypes,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012, pp. 873–882.

[7] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in Artificial Intelligence and Statistics, 2012, pp. 127–135.
[8] X. Chen, Z. Liu, and M. Sun, "A unified model for word sense representation and disambiguation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1025–1035.
[9] M. K°ageb¨ack, F. Johansson, R. Johansson, and D. Dubhashi, "Neural context embedding for automatic discovery of word senses," in Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 25–32.
[10] O. Melamud, D. McClosky, S. Patwardhan, and M. Bansal, "The role of context types and dimensionality in learning word embedding," arXiv preprint arXiv:1601.00893, 2016.
[11] Q. Liu, H. Jiang, S. Wei, Z.-H. Ling, and Y. Hu, "Learning semantic word embedding based on ordinal knowledge constraints," 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), vol. 1, 2015, pp. 1501–1511.
[12] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 238–247.
[13] Z. Haj-Yahia, A. Sieg, and L. A. Deleris, "Towards unsupervised text classification leveraging experts and word embedding," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 371–379.
[14] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embedding 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), 2019, pp. 724–728.
[15] C. Chen, S. Gao, and Z. Xing, "Mining analogical libraries in q&a discussions–incorporating relational and categorical knowledge into word embedding," in 2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER), vol. 1. IEEE, 2016, pp. 338–348.
[16] X. Yang, D. Lo, X. Xia, L. Bao, and J. Sun, "Combining word embedding with information retrieval to recommend similar bug reports," in 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2016, pp. 127–137.
[17] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli, "[journal first] sentiment polarity detection for software development," in 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE). IEEE, 2018, pp. 128–128.
[18] A. Ferrari, B. Donati, and S. Gnesi, "Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embedding," in 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW). IEEE, 2017, pp. 393–399.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," North American Association for Computational Linguistics (NAACL), 2019.
[20] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," arXiv preprint arXiv:1908.09355, 2019.
[21] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018.
[22] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in Advances in neural information processing systems, 2015, pp. 3079– 3087.
[23] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.
[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.
[25] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in Advances in Neural Information Processing Systems, 2017, pp. 6294–6305.
[26] K. Clark, M.-T. Luong, C. D. Manning, and Q. V. Le, "Semisupervised sequence modeling with cross-view training," arXiv preprint arXiv:1809.08370, 2018.
[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pretraining," URL https://s3-us-west-2. amazonaws. com/openaiassets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[28] O. Melamud, J. Goldberger, and I. Dagan, "context2vec: Learning generic context embedding with bidirectional lstm," in Proceedings of the 20th SIGNLL conference on computational natural language learning, 2016, pp. 51–61.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.
[30] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.
[31] E. Boros, R. Besanc¸on, O. Ferret, and B. Grau, "Event role extraction using domain-relevant word representations," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1852–1857.
[32] Z. Jiang, L. Li, D. Huang, and L. Jin, "Training word embedding for deep learning in biomedical text mining tasks," in 2015 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2015, pp. 625–628.
[33] X. Ye, H. Shen, X. Ma, R. Bunescu, and C. Liu, "From word embedding to document similarities for improved information retrieval in software engineering," in Proceedings of the 38th international conference on software engineering. ACM, 2016, pp. 404–415.
[34] K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera, "Synergistic union of word2vec and léxicon for domain specific semantic similarity," in 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). IEEE, 2017, pp. 1–6.
[35] B. Zhang, "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.
[36] K. Y. M. Learning, "Data science community," URl: https://www. kaggle.com.
[37] M. Choetkiertikul, H. K. Dam, T. Tran, T. T. M. Pham, A. Ghose, and T. Menzies, "A deep learning model for estimating story points," IEEE Transactions on Software Engineering, 2018.
[38] F. Dalpiaz, "Requirements data sets (user stories)," Mendeley Data, v1, 2018.
[39] T. vor der Br¨uck and M. Pouly, "Text similarity estimation based on word embedding and matrix norms for targeted marketing," 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), 2019, pp. 1827–1836.
[40] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pre-trained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.
[41] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
[42] E. M. De Bortoli Fávero, D. Casanova, and A. R. Pimentel, "SE3M: A model for software effort estimation using pre-trained embedding models," arXiv e-prints, pp. arXiv–2006, 2020.

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

AN INTELLIGENT SYSTEM TO IMPROVE ATHLETE DEPRESSION AND EATING DISORDER USING ARTIFICIAL INTELLIGENCE AND BIG DATA ANALYSIS

Xuannuo Chen[1]and Yu Sun[2]

[1]Linfield Christian, 31950 Pauba Road, Temecula, CA, 92592
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

The inspiration for the creation of this app stemmed from the deeply rooted history of eating disorders in sports, particularly in sports that emphasize appearance and muscularity which often includes gymnastics, figure skating, dance, and diving [1]. All three sports require rapid rotation in the air which automatically results in the necessity of a more stringent weight requirement. Eating disorders can also be aggravated by sports who focus on individual performances rather than team-oriented like basketball or soccer [5]. According to research, up to thirteen percent of all athletes have, or are currently suffering from a form of eating disorder such as anorexia [2] and bulimia [3]. In the National Collegiate Athletic Association, it is estimated that up to sixteen percent of male athletes and forty-five percent of female athletes have been diagnosed with an eating disorder.

KEYWORDS

Data Mining, Mobile APP, Machine Learning

1. INTRODUCTION

This app is meant for athletes to properly track their training schedule, as well as creating a proper ratio between training, eating, and resting. It allows athletes to log their daily activity level, as well as selecting their food intake and sleep hours [4]. It is applicable for athletes in all sports, as it does not contain specific restrictions. Additionally, this app contains key features such as a help page, which links all the important prevention hotlines in both English and Spanish. In creating this app, the goal is to control the levels of eating disorders in sports. Athletes are easily brainwashed by abusive coaches, and are given a falsified image of their bodies [6]. While outside factors such as these are uncontrollable, the athlete's own perception of themselves can be changed.

When first logging into the app, the sign-up screen appears first. New users have the ability to register themselves for the software, while returning users are able to log in using their email address and password. New users are asked to enter basic information about themselves, including their name and gender, which is stored in the profile page. Additionally, as an app aimed at decreasing the rates of eating disorders, athletes are asked for their current weight and height in order to determine any early signs of eating disorders.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 131-140, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111910

This app aims at reducing the rates of eating disorders through its three main components. When the athlete initially logs in, they are automatically directed to a fresh log page, where they can track their activity level, food intake, food quality, sleep hours, and sleep quality. Each category serves a specific purpose. In regards to activity level, food intake and hours of sleep should remain in a healthy ratio with their activity level. For example, two hours of sleep is insufficient for someone who has a five-hour practice period with high intensity. By entering their data, athletes are automatically promoted to reflect on their healthiness. Food intake and food quality have a stark difference fundamentally. Food intake suggests the amount of food consumed, which food quality measures the overall healthiness of the consumed food [7]. To put this into proper perspective, 600 grams of kale drastically differs from 600 grams of French fries. In addition, some foods are denser than others. 600 grams of kale can easily occupy a large salad bowl and is more than enough for one person, while 600 grams of sweet potatoes may only be one or two boiled potatoes, which only serves as a snack for many athletes, as sugar and carbohydrates are effective body fuels [8].

Similarly, sleep hours and sleep quality are vastly different. A person may lightly sleep for nine hours, and therefore feels fatigued the following day. Other times, athletes may sleep for only six hours, but become extremely energetic. Not only does the amount of sleep needed depend on the sport, but it also depends on the athlete. If only sleep hours were implemented, it would not be enough to personally measure the healthiness level of athletes. This is to prevent athletes from being pressured to over-sleeping, as it can result in unexpected tiredness. Normally, scientific research tends to suggest eight hours of sleep for young adults, while many may only need six hours.

Once the athlete submits their data for the day, they are then directed to the ‘Dashboard’ page, which tracks all of their past data. Users are permitted to access their past information by clicking on the tile. Having the ability to view past history allows the athletes to understand whether their eating and sleeping habits have improved. The athletes who consistently rank themselves as poor in regards to sleep quality rating can be identified as those who need sleep improvements. Seeing this, athletes can thereby apply for necessary help, including sleep treatments. Similarly, athletes can track their eating routine in proportion to their training hours and sleep levels. Oftentimes, those who deprive themselves of food do not have enough body energy to fuel for an entire session of practice. This can result in nausea, headaches, weak bodies, and in severe cases, injuries or fainting [9]. A special feature on the dashboard page is the randomly generated inspirational quotes from prominent athletes. These include Simone Biles, Michael Phelps, and many others. The purpose of this is simply to encourage struggling athletes on their path to success.

The last component of the app is the ‘Help’ page. In this page, athletes have access to a number of resources including websites and phone numbers to different hotlines. The National Suicide Hotline is placed on top in both English and Spanish as it is the most urgent and important. In the case of an emergency, athletes or those around them can rapidly access a series of phone numbers.

In general, eating disorders are directly related to abusive coaching methods. Many coaches have a tendency to blame the losses of athletes on their weight. A prime example of this is the experience of track runner Mary Cain. At seventeen years of age, Cain was one of the fastest runners in the United States, and qualified for the 2013 World Championships team as the youngest American in history. In the same year, Cain was signed to Nike’s Oregon Project, the leading track and field program at the time, which was led by coach Alberto Salazar. As a coach who was only familiar with male athletes, Salazar became obsessed with Cain’s weight, and revolved the entire training schedule on reducing weight. As a result of her consistent malnutrition, Cain’s body began breaking down. Her bones had become fragile [15], which began

to break easily, and she had also developed serious depression which resulted in actions of self-harm [10].

The abusive methods of coaching cannot be controlled by the opinions of the general public, as most of them are merely result-driven. Regulation should come from major organizations such as Safesport. The goal of this app is to manage what can be controlled, which is the athlete's perception of themselves. Through publishing this app, the hope is to see a reduction in the symptoms of eating disorders in users. This can be verified through changes made in the 'log' page. Improvements are justified through a change in behaviour, for example, consuming more healthy food and getting better sleep quality.

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 related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Finding an Adequate Solution

One major challenge that occurred throughout the process was finding an adequate solution to the original problem that was posed. Eating disorders are extremely overlooked in sports, meaning that minimal research has been done, and almost no preventative measures have been taken, especially in performance-based sports. In fact, it is basically impossible to eliminate eating disorders through one app, as the environment an athlete is in is the most outcome-defining factor. The most an app can do is to limit the potential of further developing eating disorders.

2.2. Naming Each Label

Additionally, it was difficult to find a proper wording for every label in the app. It was necessary to ensure that the words being used are not triggering under any circumstances. Requesting the user of the app to enter their weight upon registering can already cause insecurities among those who struggle with their body-image. Therefore, the selection of words must be done carefully and precisely. Words such as "fat" or "overweight" must be completely abandoned. This is because those who struggle with eating-disorders do not have an accurate perception of body-weight. Despite being possibly underweight, they may still believe the opposite, most likely due to the manipulative environment they practice in.

2.3. Asking Athlete to Fill Logs

In the end, it was determined that the most helpful method is asking their athlete to fill out daily logs that tracks their progress. This ensures that the app respects the user's privacy and sensitivity while promoting a healthier diet culture. Athletes can view their own progress by reviewing their past entries. For example, improvements are shown if the user historically indicates that they do not eat enough, and after two weeks, they start to eat more.

3. SOLUTION

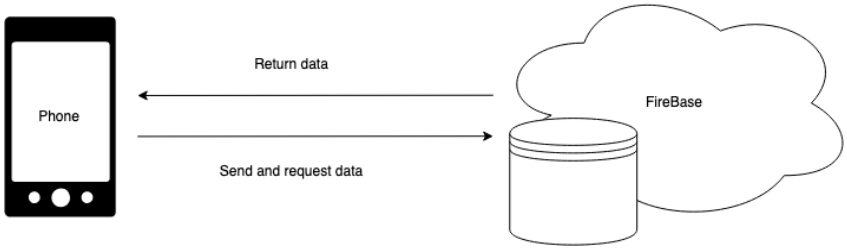


Figure 1. The overview of the project

Affly is a food health tracking app that helps users track the quality and quantity of their food and their sleeping quality. This is done by asking the user to input how much they eat and the quality of their food from a drop down based on how they feel. This subjective method of measuring and rating allows the user to reflect, think about and have an honest discussion with themselves about their habits.This method also avoids the difficulty of counting calories and tracking macro and micro nutrition [12]. Users are encouraged to add new data points daily and once the app has enough data it can show the users' habits over time and give users insights.

Affly utilities Google's Firebase tools to manage its users and their data. User creation is done within the Affly app itself only requiring an email for registration using Firebase's Authentication services. User data is stored using Firebase's Realtime Database. The information is stored using the users' unique ID and stores daily data using timestamps as keys under the users' unique ID [13].
Afflywas created using the flutter framework to allow easy development for both iOS and Android [14].

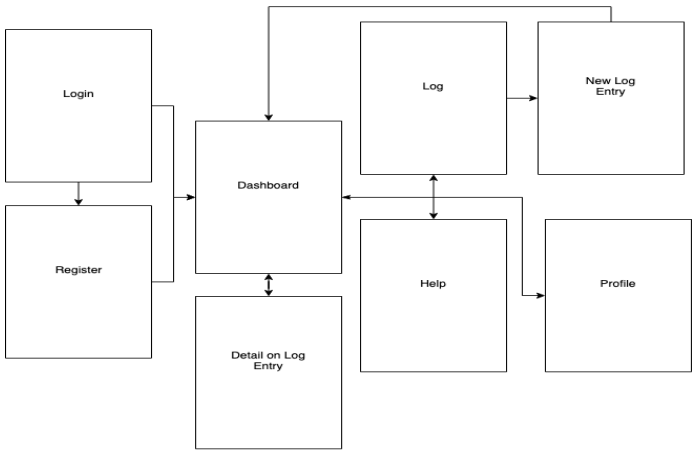


Figure 2. Basic process

Login:Allows the user to login.
Register: Allows the user to sign up and register.
Dashboard: Shows the user the latest insights of their health.
Log: Shows a list of log entries the user has entered. Pressing on a log entry lead to a detail page on the log entry.

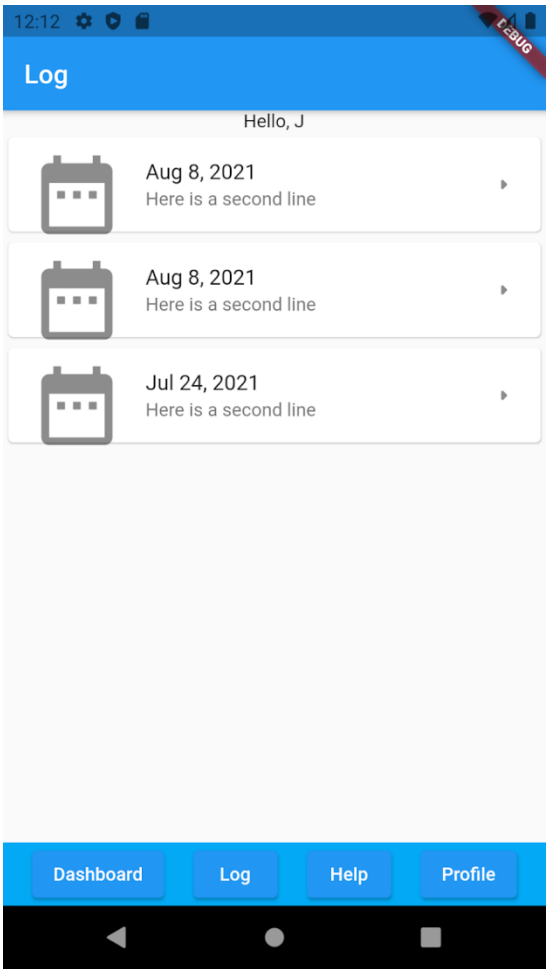


Figure 3. Screenshot of the App (1)

Log Entry Detail Page: Shows eatil about the log entry.
New Log Entry: This page allows the user to add a new entry.

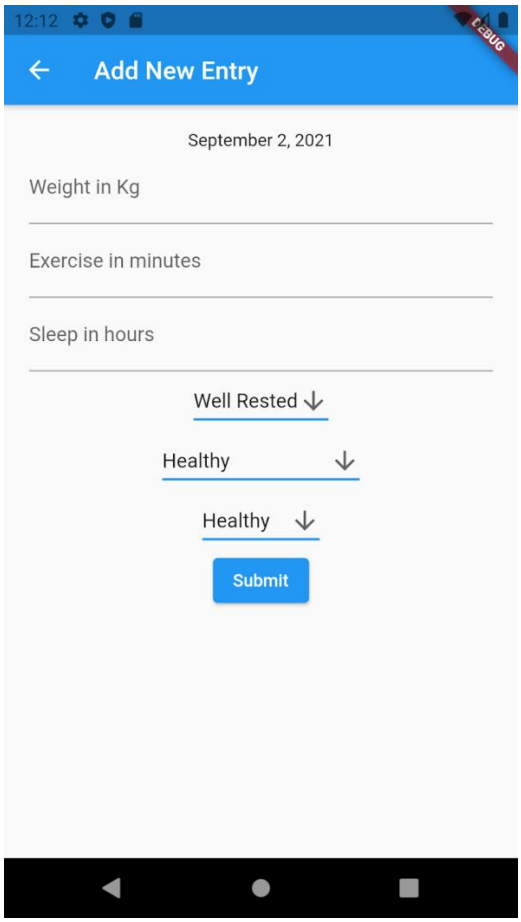


Figure 4. Screenshot of the App (2)

Help: This page lists resource users can contact for support.
Profile: This page shows the user info such as their email and profile name.

4. EXPERIMENT

4.1 Experiment 1

To prove the app works effectively and efficiently in the outdoor sports area, we ask 100 high school students who like to do sports like camping, walking trails, climbing mountains, and other outdoor sports, the Table shows the scores given by different types of sports lovers and the charts shows the ratio and distribution. We tested with 4 different types of sports and the test group is from different areas of Los Angeles so the group amount is big enough and group Diversity is guaranteed.

	Total	Very Helpful	Helpful	Normal	Not help
camping	25	20	3	1	1
walking trails	25	19	4	2	0
climbing mountains	25	22	2	0	1
other outdoor sports	25	20	5	0	0

Figure 5. Result Matrix

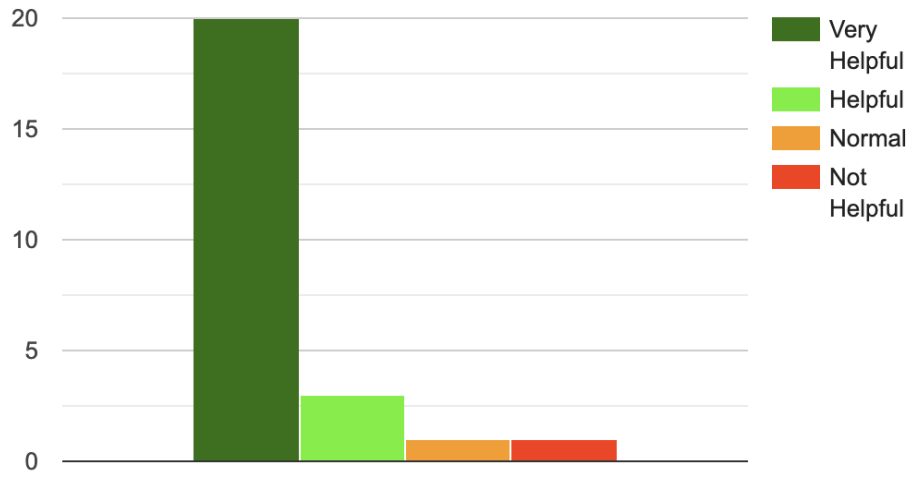


Figure 6. Score by Camping lover ratio

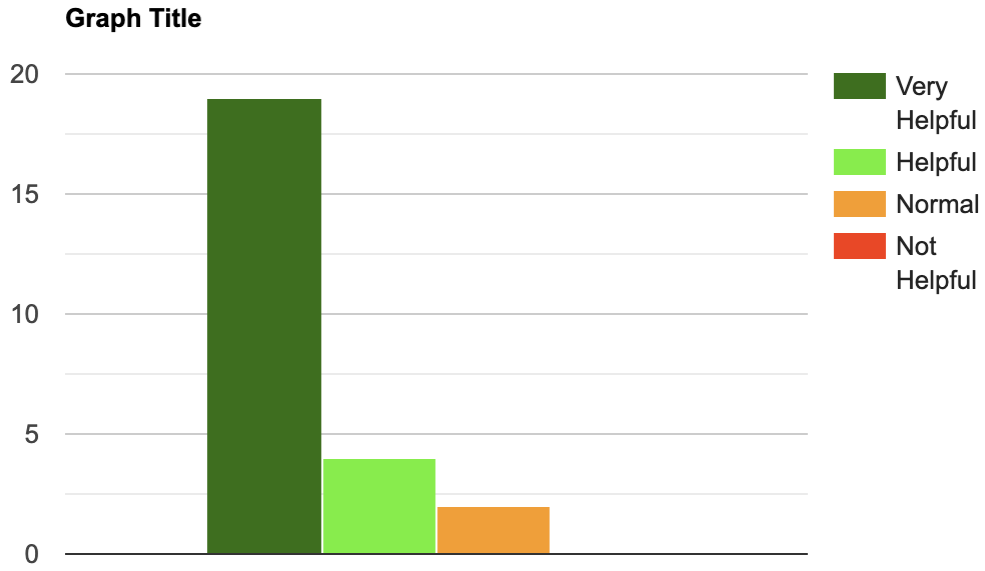


Figure 7. Score by walking trails ratio

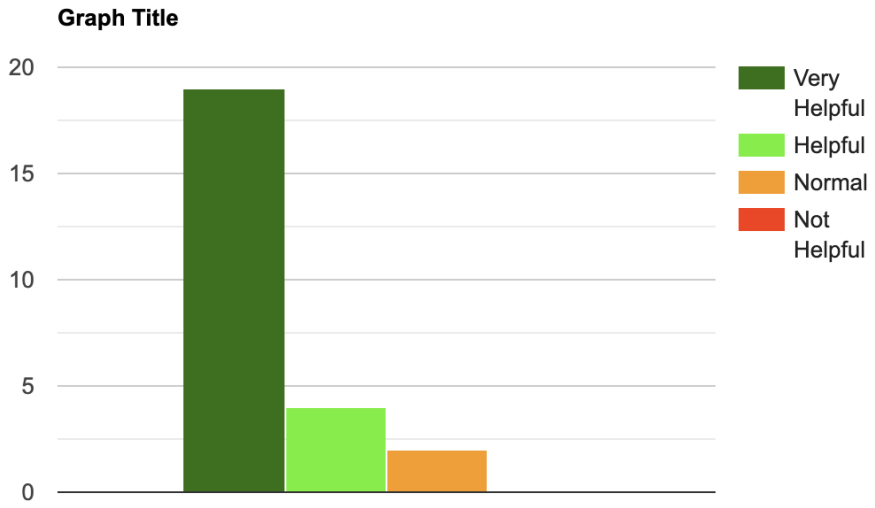


Figure 8. Score by climbing mountains ratio

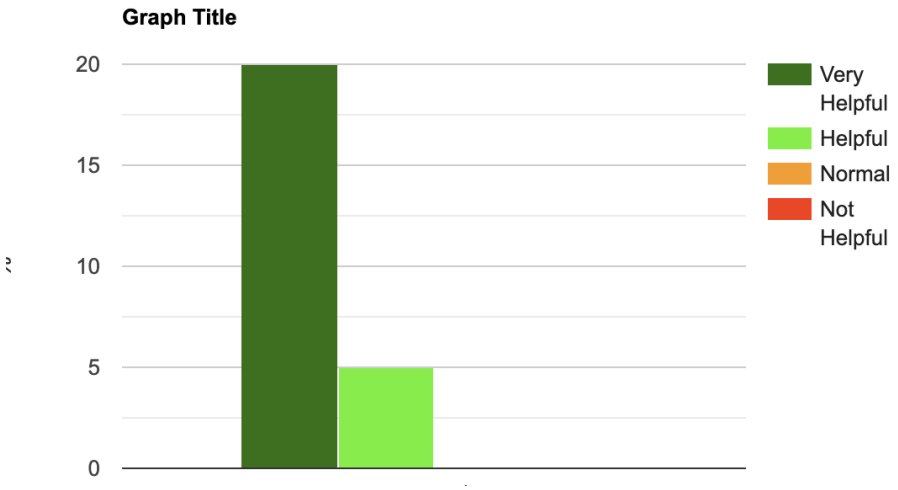


Figure 9. Score by other sports ratio

4.2. Experiment 2

For the second survey, we find 50 people who are working on training muscles and divide them into two groups. One group uses the app and the other one never uses it. We calculate the changes in the body fat rate of both groups. The two groups are following the same workout schedule and plan so It turns out people who work with the app reduce their body fat rate much faster than people who work without the app. The group works with the app to reduce their body rate average by 0.8%, and the group working without the app reduces their body rate average by 0.3%.

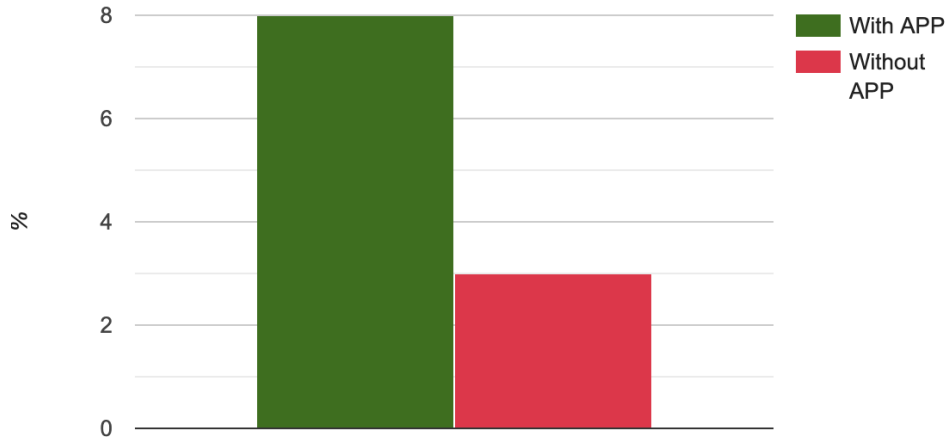


Figure 10. Survey result

Based on the above two experiments, we can prove that people who work with the app reduce the body fat rate much faster than people who work without the app. So we can consider the app to be effective and efficient.

5. RELATED WORK

A majority of last research done on eating disorders focuses on the treatments of eating disorders rather than the detection of earlier signs. This is because for a computer program, it is much easier to function through regulation. Most treatments will be similar for all users. On the other hand, early symptoms of eating disorders can vary, which makes it extra difficult to create an app that can accurately predict eating disorders.

6. CONCLUSIONS

The main component of this project is the construction of the application itself. To begin with, four parts of the app were proposed, including the “dashboard”, “log”, “help”, and “profile” pages. The “log” and “help” pages are the most functional and crucial. In contrast, “dashboard” and “profile” are used for display purposes. Users will likely spend the longest time on “log” as they need to consistently track their daily progress. The dashboard indicates past entries and progress the user has made. Under normal circumstances, the user should not be constantly accessing the “help” page, however, if necessary, the “help” page is a valuable resource that can be potentially life-saving. Tests have been performed to test the overall functionality of the app. The “sign-up” and “log-in” pages are able to effortlessly upload users' information to the database to ensure that their history on the app does not become erased. The “log” page is able to accurately record the user’s daily entries which are displayed in chronological order in the “dashboard”.

While the app functions effectively and efficiently, there is plenty of room left for improvement. In the future, this app seeks to include automatic notifications. After tracking and analyzing all of the user’s information, the app will automatically generate reports which will be returned to the user at their convenience. The user will have access to the app in the dashboard and will no longer be required to run their own analysis. This would be able to solve the app’s largest issue: practicability [11]. Users who are already struggling with eating disorders will not have the time and energy to read about their past eating history.

REFERENCES

[1] Hodge, Archibald, and Benjamin B. Warfield. Inspiration. Wipf and Stock Publishers, 2008.
[2] Bell, Rudolph M. Holy anorexia. University of Chicago Press, 2014.
[3] Pyle, Richard L., James E. Mitchell, and Elke D. Eckert. "Bulimia: a report of 34 cases." The Journal of Clinical Psychiatry (1981).
[4] Wardle, Jane, Kathryn Parmenter, and Jo Waller. "Nutrition knowledge and food intake." Appetite 34.3 (2000): 269-275.
[5] Polivy, Janet, and C. Peter Herman. "Causes of eating disorders." Annual review of psychology 53.1 (2002): 187-213.
[6] Satel, Sally, and Scott O. Lilienfeld. Brainwashed: The seductive appeal of mindless neuroscience. Basic Civitas Books, 2013.
[7] Kriss, Max. "THE FOOD CONSUMED, THE HEAT PRODUCTION, THE." Journal of Agricultural Research 40 (1930): 283.
[8] Burke, Louise M., et al. "Carbohydrates for training and competition." Journal of sports sciences 29.sup1 (2011): S17-S27.
[9] Stern, Robert Morris, Kenneth L. Koch, and Paul Andrews. Nausea: mechanisms and management. OUP USA, 2011.
[10] Pirlich, Matthias, et al. "The German hospital malnutrition study." Clinical nutrition 25.4 (2006): 563-572.
[11] Gioia, Dennis A. "Practicability, paradigms, and problems in stakeholder theorizing." Academy of Management Review 24.2 (1999): 228-232.
[12] Shenkin, Alan. "The key role of micronutrients." Clinical nutrition 25.1 (2006): 1-13.
[13] Klyne, Graham, and Chris Newman. Date and time on the internet: Timestamps. RFC 3339, July, 2002.
[14] Leshem, Shosh, and Vernon Trafford. "Overlooking the conceptual framework." Innovations in education and Teaching International 44.1 (2007): 93-105.
[15] Osaghae, Eghosa E. "Fragile states." Development in Practice 17.4-5 (2007): 691-699.

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

THE USE OF BIG DATA IN MACHINE LEARNING ALGORITHM

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China

ABSTRACT

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. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Such minimal human intervention can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyse some of the different machine learning algorithms and methods which can be applied to big data analysis, as well as the opportunities provided by the application of big data analytics in various decision making domains.

KEYWORDS

Artificial Intelligence, Big Data Analysis, Machine Learning.

1. INTRODUCTION

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyse bigger, more complex data and deliver faster, more accurate results – even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks [1].

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum. While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data, over and over, faster and faster – is a recent development [2]. This paper will look at some of the different machine learning algorithms and methods which can be applied to big data analysis, as well as the opportunities provided by the application of big data analytics in various decision making domains.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 141-147, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111911

2. HOW MACHINE LEARNING WORKS

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy [3].

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics [4]. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

2.1. Machine Learning Algorithms

Machine learning algorithms can be categorize into three main parts:

1. **A Decision Process**: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabelled, your algorithm will produce an estimate about a pattern in the data.
2. **An Error Function**: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.
3. **A Model Optimization Process**: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

2.2. Types of Machine Learning Methods

Machine learning classifiers fall into three primary categories [5]:

2.2.1. Supervised machine learning

Supervised learning also known as supervised machine learning, is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

2.2.2. Unsupervised machine learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, crossselling strategies, customer segmentation, image and pattern recognition. It’s also used to reduce the

number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, kmeans clustering, probabilistic clustering methods, and more [6].

2.2.3. Semi-supervised learning

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labelled data set to guide classification and feature extraction from a larger, unlabelled data set. Semi-supervised learning can solve the problem of having not enough labelled data (or not being able to afford to label enough data) to train a supervised learning algorithm.

2.3. Practical use of Machine Learning

Here are just a few examples of machine learning you might encounter every day [7]:

2.3.1. Speech Recognition

It is also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, and it is a capability which uses natural language processing (NLP) to process human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.

2.3.2. Customer Service

Online chatbots are replacing human agents along the customer journey. They answer frequently asked questions (FAQs) around topics, like shipping, or provide personalized advice, cross-selling products or suggesting sizes for users, changing the way we think about customer engagement across websites and social media platforms. Examples include messaging bots on e-commerce sites with virtual agents, messaging apps, such as Slack and Facebook Messenger, and tasks usually done by virtual assistants and voice assistants.

2.3.3. Computer Vision

This AI technology enables computers and systems to derive meaningful information from digital images, videos and other visual inputs, and based on those inputs, it can take action. This ability to provide recommendations distinguishes it from image recognition tasks. Powered by convolutional neural networks, computer vision has applications within photo tagging in social media, radiology imaging in healthcare, and self-driving cars within the automotive industry.

2.3.4. Recommendation Engines

Using past consumption behaviour data, AI algorithms can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

2.3.5. Automated stock trading

Designed to optimize stock portfolios, AI-driven high-frequency trading platforms make thousands or even millions of trades per day without human intervention.

3. WHAT IS BIG DATA AND WHAT ARE ITS BENEFITS

Big data analytics has revolutionized the field of IT, enhancing and adding added advantage to organizations. It involves the use of analytics, new age tech like machine learning, mining, statistics and more. Big data can help organizations and teams to perform multiple operations on a single platform, store Tbs of data, pre-process it, analyse all the data, irrespective of the size and type, and visualize it too [8].

The Sources of Big Data:

Black Box Data

This is the data generated by airplanes, including jets and helicopters. Black box data includes flight crew voices, microphone recordings, and aircraft performance information.

Social Media Data

This is data developed by such social media sites as Twitter, Facebook, Instagram, Pinterest, and Google+.

Stock Exchange Data

This is data from stock exchanges about the share selling and buying decisions made by customers.

Power Grid Data

This is data from power grids. It holds information on particular nodes, such as usage information.

Transport Data

This includes possible capacity, vehicle model, availability, and distance covered by a vehicle.

Search Engine Data

This is one of the most significant sources of big data. Search engines have vast databases where they get their data.

The speed at which data is streamed, nowadays, is unprecedented, making it difficult to deal with it in a timely fashion. Smart metering, sensors, and RFID tags make it necessary to deal with data torrents in almost real-time. Most organizations are finding it difficult to react to data quickly. Not many years ago, having too much data was simply a storage issue [9]. However, with increased storage capacities and reduced storage costs are now focusing on how relevant data can create value.

There is a greater variety of data today than there was a few years ago. Data is broadly classified as structured data (relational data), semi-structured data (data in the form of XML sheets), and unstructured data (media logs and data in the form of PDF, Word, and Text files). Many companies have to grapple with governing, managing, and merging the different data varieties [10].

3.1. Advantages of Big Data

1. Today's consumer is very demanding. All customer wants to be treated as an individual and to be thanked after buying a product. With big data, supplier will get actionable data that they can use to engage with their customers one-on-one in real-time [11]. One way big data allows supplier to do this is that they will be able to check a complaining customer's profile in real-time and get info on the product(s) the customer is complaining about. Supplier will then be able to perform reputation management.
2. Big data allows supplier to re-develop the products/services they are selling. Information on what others think about their products, such as through unstructured social networking site text helps supplier in product development.
3. Big data allows supplier to test different variations of CAD (computer-aided design) images to determine how minor changes affect their process or product. This makes big data invaluable in the manufacturing process.
4. Predictive analysis will keep supplier ahead of their competitors. Big data can facilitate this by, as an example, scanning and analysing social media feeds and newspaper reports. Big data also helps supplier do health-tests on their customers, suppliers, and other stakeholders to help supplier reduce risks such as default.
5. Big data is helpful in keeping data safe. Big data tools help supplier map the data landscape of their company, which helps in the analysis of internal threats. As an example, supplier will know if their sensitive information has protection or not. A more specific example is that supplier will be able to flag the emailing or storage of 16 digit numbers (which could, potentially, be credit card numbers) [12].
6. Big data allows supplier to diversify their revenue streams. Analysing big data can give supplier trend-data that could help the supplier come up with a completely new revenue stream.
7. The supplier website needs to be dynamic if it is to compete favourably in the crowded online space. Analysis of big data helps supplier personalize the look/content and feel of their site to suit every visitor based on, for example, nationality and sex. An example of this is Amazon's IBCF (item-based collaborative filtering) that drives its "People you may know" and "Frequently bought together" features [13].
8. If the supplier is running a factory, big data is important because the supplier will not have to replace pieces of technology based on the number of months or years they have been in use. This is costly and impractical since different parts wear at different rates. Big data allows supplier to spot failing devices and will predict when the supplier should replace them.
9. Big data is important in the healthcare industry, which is one of the last few industries still stuck with a generalized, conventional approach. Big data allows a cancer patient to get medication that is developed based on his/her genes.

3.2. Challenging of Big Data

1. One of the issues with big data is the exponential growth of raw data. The data centres and databases store huge amounts of data, which is still rapidly growing. With the exponential growth of data, organizations often find it difficult to rightly store this data [14].
2. The next challenge is choosing the right big data tool. There are various big data tools, however choosing the wrong one can result in wasted effort, time and money too.
3. Next challenge of big data is securing it. Often organizations are too busy understanding and analysing the data, that they leave the data security for a later stage, and unprotected data ultimately becomes the breeding ground for the hackers.

4. CONCLUSIONS

So this study was concerned by understanding the inter-relationship between machine learning and big data analysis, what frameworks and systems that worked, and how machine learning can impact the big data analytic process whether by introducing new innovations that foster advanced machine learning process and escalating power consumption, security issues and replacing human in workplaces. The advanced big data analytics and machine learning algorithms with various applications show promising results in artificial intelligence development and further evaluation and research using machine learning are in progress.

REFERENCES

[1] Shi, Z., (2019). Cognitive Machine Learning. *International Journal of Intelligence Science*, 9, pp. 111-121.
[2] Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B., (2015). Human-Level Concept Learning through Probabilistic Program Induction. Science, 350, pp. 1332-1338.
[3] Silver, D., Huang, A., Maddison, C.J., et al., (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 529, pp. 484-489.
[4] Fukushima, K., (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36, pp. 193-202.
[5] Lecun, Y., Bottou, L., Orr, G.B., et al., (1998). Efficient Backprop. Neural Networks Tricks of the Trade, 1524, pp. 9-50.
[6] McClelland, J.L., et al., (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. Psychological Review, 102, pp. 419-457.
[7] Kumaran, D., Hassabis, D. and McClelland, J.L., (2016). What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends in Cognitive Sciences, 20, pp. 512-534.
[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] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, (2013). A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, pp. 17-22.
[11] Z. Huang, (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.
[12] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, (2015). Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79, pp. 3-15.
[13] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, (2014). The rise of big data on cloud computing: Review and open research issues, Information Systems, 47, pp. 98-115.
[14] L. Wang and J. Shen, (2013). Bioinspired cost-effective access to big data, International Symposium for Next Generation Infrastructure, pp. 1-7.

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.

DATA AUGMENTATION AND TRANSFER LEARNING APPROACHES APPLIED TO FACIAL EXPRESSIONS RECOGNITION

Enrico Randellini, Leonardo Rigutini and Claudio Saccà

QuestIT Research Lab, Siena (Italy)

ABSTRACT

The face expression is the first thing we pay attention to when we want to understand a person's state of mind. Thus, the ability to recognize facial expressions in an automatic way is a very interesting research field. In this paper, because the small size of available training datasets, we propose a novel data augmentation technique that improves the performances in the recognition task. We apply geometrical transformations and build from scratch GAN models able to generate new synthetic images for each emotion type. Thus, on the augmented datasets we fine tune pretrained convolutional neural networks with different architectures. To measure the generalization ability of the models, we apply extra-database protocol approach, namely we train models on the augmented versions of training dataset and test them on two different databases. The combination of these techniques allows to reach average accuracy values of the order of 85% for the InceptionResNetV2 model.

KEYWORDS

Computer Vision, Facial Recognition, Data Augmentation, Transfer Learning.

1. INTRODUCTION

The ability to build intelligent systems that accurately recognize the emotions felt by a person is an open challenge of Artificial Intelligence and undoubtedly represents one of the points of contact between the human and machine spheres. Since the face expression is the first thing we pay attention to when we want to understand a person's state of mind, facial expression analysis represents the first step in researching and building a human emotion classifier. In the facial expression recognition (FER) task, it is believed that there are six basic universal expressions, namely fear, sad, angry, disgust, surprise and happy [1]. To these emotions is often added a neutral expression.

Thanks to recent advances in the field of Machine Learning and Deep Learning, many FER systems have been proposed in the literature over the years, obtaining in some cases high accuracy values [2] [3]. On the other hand, greater levels of precision can be achieved taking into account the following issues:

1) it is observed a significant overlap between basic emotion classes [1] and differences in cultural manifestation of a given emotion [4];
2) the public image-labeled databases widely used to train and test FER systems may not be large enough;
3) the available datasets differ in the quality of pictures and how people express a given emotion. Some of databases are composed of images taken 'in the wild', where the labeled

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 149-163, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111912

emotion is naturally manifested by the people while doing some action. This differs from other datasets where the pictures are taken when the people are posing with that particular expression;

4) the results strongly depend on the databases used for train and test the models. In the intra-database protocol, where train is carried out in one database and test in a subject independent set of the same database, the current methods achieve high accuracy, reaching around 95% [5] [6]. On the contrary, methods evaluated in the cross-database protocol, where train is carried out in one or more databases and the test in different databases, usually are obtained lower accuracy, ranging between 40% and 88% [5] [6] [7] [8].

An automated FER system can be seen as a supervised classification method comparing selected facial features from given image or video frame with faces within a database. It is a well established fact that computer vision tasks are optimally solved by convolutional neural network (CNN) and, it is usually necessary to have large databases in order to avoid overfitting [11][12][13]. Unfortunately, some public image-labeled databases used to train and test FER systems, such as Karolinska Directed Emotional Faces (KDEF)[14] and Extended Chon-Kanade (CK+)[15], are not sufficiently large. To overcome this problem were introduced data augmentation (DA) techniques. They are of two types: (i) geometric (e.g. rotation, translation and scaling) and color transformations that change the shape or the color of the starting images leaving unchanged their labels [16][7][8]; (ii) guided-augmentation methods (e.g. by generative adversarial network (GAN) [17]) that create new synthetic images with specific labels [18][8].

Another way to circumvent the obstacle of small train databases is making use of transfer learning and fine tuning. These are machine learning techniques enabling to use knowledge from previously learned tasks and apply them to newer, related ones [19][11].

Leveraging on the previous techniques of data augmentation and transfer learning, the recent work of Zavarez et al.[7] proposed a cross-database evaluation where a pre-trained VGG16 network is fine tuned on six databases, augmented by using geometrical transformations, and evaluated on a seven different database. Their test on CK+ database reaches an average accuracy of 88%. In a similar way, Porcu et al.[8], augmenting the train database KDEF with synthetics images by means of geometrical transformations and GAN techniques, reach an accuracy of 83% when a pre-trained VGG16 neural network is evaluated always in the CK+ test set.

The aim of this paper is to explore whether it is possible to further improve the accuracy and the ability to generalize on new data of automated FER systems. We will examine if the available data augmentation techniques allow us to enlarge the training datasets more than what has been done so far. Moreover, we will consider different CNN architectures in addition to the already used VGG16.

To address the issues related to the small size of KDEF database, we will make use of both DA techniques exposed above. We will apply geometric and color transformations in an offline mode storing the results as a new database. After that, we will build GAN models from scratch to generate novel synthetic images for every emotion. Moreover, in order to compare the results and enlarge the training dataset even further, we will make use of the synthetic images kindly made available by the group of Porcu et al.[8]. Our results will show that as the number of training data increases, will improve also the stability and performances of the models.

Inspired by the previous works [7] and [8], in this paper we will conduct both cross-database and intra-database protocol experiments. Once we have trained the models on the full KDEF dataset and its enlarged versions, we will evaluate them on the CK+ and JAFFE test set, showing a good ability to generalize on new data. Furthermore, we will apply a k-fold cross validation making

use of a general database obtained by the union of the KDEF dataset, its augmented versions, and the CK+ and JAFFE datasets.

Encouraged by the remarkable results obtained in the field of image recognition, in this paper we will make use of transfer learning techniques applied to other CNN architectures not used before. In addition to the VGG16, we will consider the VGG19 [21], InceptionV3 [22] and InceptionResNetV2 [23] architectures already pre-trained on the ImageNet dataset [20]. Then, simply by modifying the finals layers of each models and fine tuning their values along the KDEF dataset and its augmented versions, we will be able to reach high accuracy values for the problem of face emotion recognition.

For example, we will show that the InceptionResNetV2 network applied to the CK+ dataset reaches a mean accuracy of 86.15% with a very close range variability. This is an index of excellent stability and generalization to new data.

The work is structured as follow. In section 2 we will describe the three datasets selected for the comparison (KDEF, CK+ and JAFFE) and as we pre-processed the images. In section 3, we will illustrate how we have increased the data by means of geometric transformations and GAN techniques and build the different train sets. In section 4, we describe how we have applied transfer learning and fine tuning techniques on pre-trained CNN. The section 5 describes the experimental setup and reports the results. We conduct our experiments by using Python 3.7.10, Tensorflow 2.4.1, and 12GB NVIDIA Tesla K80 GPU. Finally, in section 6, we present the conclusions and remarks.

2. DATA PREPARATION

2.1. Dataset

We conduct our analysis making use of three databases of images from subjects of different ethnicities, genders, and ages in a variety of environments: KDEF, CK+ and JAFFE databases. Their main properties are reported in the following:

1) KDEF: The Karolinska Directed Emotional Faces (KDEF) [14] consists of 4900 pictures of 70 subjects (35 males and 35 females), each of which has been photographed twice in each of the seven facial expressions at five different angles (full left profile, half left profile, straight, half right profile, full right profile). For our experiments we consider only the straight images with a total of 980 pictures.
2) CK+: The Extended Cohn-Kanade (CK+) [15] consists of 100 university students aged from 18 to 30 years. Each picture is a frame from videos where each subject was instructed to perform expressions that begin and end with the neutral expression. Once neglected the images belonging to the contempt expressions, which is not included in the list of considered emotions, we get a total of 902 pictures.
3) JAFFE: The JAFFE dataset [24] consists of 213 images of different facial expressions from 10 different Japanese female subjects. Each subject was asked to do seven facial expressions (six basic facial expressions and neutral).

In order to train a supervised classifier in the cross-database protocol, we take the KDEF as starting point to build the final training databases. Namely, on the KDEF we will apply various data augmentation techniques to obtain four different enlarged training databases. Finally, the models will be tested on the CK+ and JAFFE.

2.2. Face detection and image standardization

To understand what kind of emotion a person is feeling, we look at his eyes, if he wrinkles his nose, the shape of mouth and so on. All of these features manifest on the face of the person we are looking at, thus we have to concentrate only on the face, neglecting other parts of the body and the background. Our first action is to reduce all the images just to the rectangle containing the face. We run this transformation using the DNN module of the OpenCV library [25], with a confidence of 0.5 for face recognition.

We also fix a standard dimension for the input images, now containing only the portion of the face. We adopt (224, 224, 3), where the first two dimensions represent the number of the row and column pixels, while the third dimension is the number of colour channels in the RGB sequence. At this level we leave the pixel intensity between 0 and 255. As better explained in Section (4.3), we will change the normalization of the input values depending on the classifier model.

3. DATA AUGMENTATION

The KDEF is a small train database to solve a complex task of computer vision thus, in order to increase the amount of training data, we perform a Data Augmentation step. In particular, to generate new data from existing ones, we follow two approaches:

1) Geometrical and colour transformations;
2) Generation of synthetic images from scratch using GAN

3.1. Geometrical and colour transformations

In the first approach, we build a set of artificial synthetic images by modifying some geometrical and color characteristics of the original images. Namely, we define the following set of operators acting on the geometry and colors of each image leaving unchanged the expression of the face:

1) Random rotation: a function that rotates each image by a random factor ρ, namely a float which denotes the upper limit, as a fraction of 2π, for clockwise and counterclockwise rotations. In the experiments we set $\rho = 0.1$.
2) Random zoom: a function which randomly zoom each image by a random factor ζ. In the experiments we set $\zeta = 0.1$.
3) Random flip: a function which randomly flip each image on the horizontal mode.
4) Random height: a function which randomly adjusts the height by a random factor θ, namely a positive float representing lower and upper bound for resizing the image vertically. In the experiments we set $\theta = 0.2$.
5) Random width: a function which randomly adjusts the width by a random factor ω, namely a positive float representing lower and upper bound for resizing the image horizontally. In the experiments we set $\omega = 0.2$.
6) Random contrast: a function which randomly adjust the contrast of an image between $[1 - \gamma, 1 + \gamma]$. In the experiments we set $\gamma = 0.2$.

We apply the above transformations five times to each original image, obtaining five synthetic images with the same target emotion. In this way, we obtain a new dataset, called KDEF_DA_OL (standing for Data Augmentation Offline) made of 4900 new images. Finally, merging KDEF_DA_OL with the starting KDEF dataset, we get the first training dataset made of 5880 images. We call this dataset KDEF_OL.

3.2. Synthetic data generation: GAN

Introduced in 2014 [17], Generative Adversarial Network (GAN) are able to learn how to reproduce synthetic data that looks real. For example, computers can learn how to create realistic images and pictures of peoples that do not exist in reality. Generally, GANs train two neural networks simultaneously: the generator attempts to produce a realistic image to fool the discriminator, which tries to distinguish whether this image is from the training set or the generated set.

The authors of [8] used the GAN framework implemented for the DeepFake autoencoder architecture of the FaceSwap project (https://github.com/deepfakes/faceswap). Basically, the face images from the KDEF database are used as base to create novel synthetic images using the facial features of two images selected from the YouTube-Faces database [26]. The novel images differ between each other, in particular with respect to the eyes, nose and mouth, whose characteristics are taken from the two selected new images. The authors kindly shared their augmented database with 980 synthetic images that, for convenience, we call KDEF_GAN_PFA. Once standardized, we merge this set of images with the previous KDEF_OL dataset in order to obtain a larger dataset with 6860 images including the original KDEF and the augmented version with both offline and GAN techniques. We call this dataset KDEF_PFA.

In this work we apply GANs models for data augmentation in a different way than [8]. First, we group the pictures with the same expression of the KDEF database. To each group we add four images of famous actors with the same expression taken from the web. For example, the pictures in Figure 1 with a manifestly happy expression, once standardized, have been added to the 140 happy images of KDEF dataset. Thus, for each emotion we obtain a set of 144 pictures that will be used to train a couple of GANs generator and discriminator networks in order to generate new synthetic images sharing the same facial expression. We believe that this procedure introduces a certain variability into the train dataset, thus the produced synthetic faces will slightly differ from the parent faces in terms of pose, brightness and background.

Figure 1. Pictures of known actors with a happy facial expression taken from the web

We assemble the discriminator model as a typical image classifier. In agreement with the structure of deep convolutional GAN (DCGAN) [27], as represented in Figure 2 the discriminator is a network made of a first convolutional 2D layer followed by five convolutional layers with striding to downscale the image by a factor of two every step. The result goes through flatten layer, followed by a dense sigmoid layer which returns a single output probability to classify the input picture as real or fake. In each striding layer we use a LeakyReLU activation function and a number of filters starting from 32 and doubling at each layer.

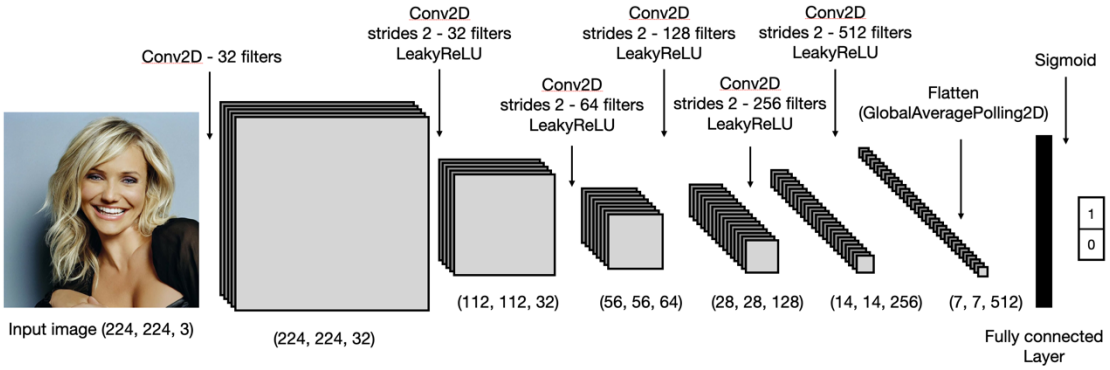


Figure 2. Discriminator architecture

As shown in Figure 3, the generator takes in input a noise vector from the latent dimension, chosen equal to 100, and generates an image. The network first upsamples the noise vector with a dense layer in order to have enough values to reshape into the first generator block. Each block consists of a transposed convolution 2D layer to upsample the image by a factor of two. We use 5 decoder blocks with LeakyReLU activation function and a final convolution 2D layer with hyperbolic tangent activation function to get a 3D tensor with the desired shape (224, 224, 3), which represents the final produced image.

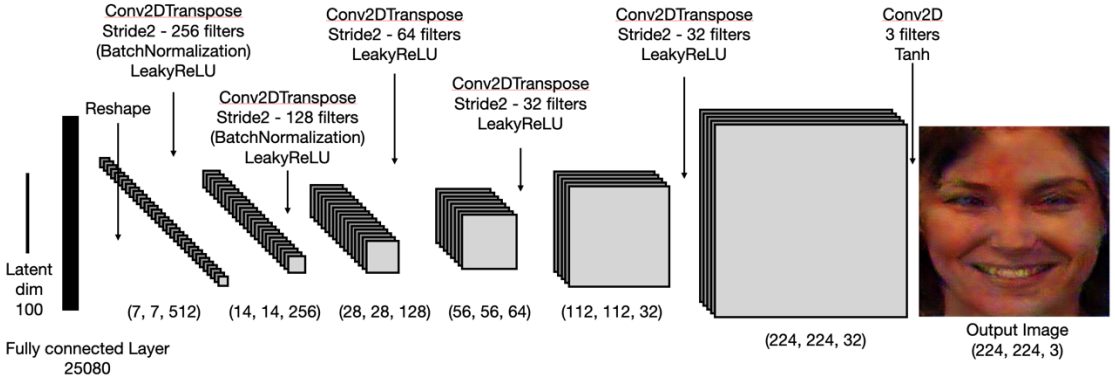


Figure 3. Generator architecture

We build the joined DCGAN by adding the discriminator on the top of the generator and train it applying the following steps. First, we send random noise to the generator, adding the output with real images to initially train only the discriminator. Then we freeze the discriminator and train the generator with the purpose to fooling the discriminator. We repeat this process iteratively for 2000 epochs using the Adam optimizer with a learning rate of 0.0002 and $\beta_1 = 0.5$ [28], and monitoring the quality of the generated images every 100 epochs. We start saving trained models starting with 1000-th epoch in order to use the most stable version depending on the quality of the produced images.

We repeat this procedure for each emotion. First, we train a DCGAN model, thus, using the most stable version, we generate 150 synthetic images. At the end we get with a total of 1050 fake images composing the dataset KDEF_GAN_Q. As before, we merge this dataset together the previous KDEF_OL in order to obtain another larger dataset with 6930 images including the

original one and the augmented version with both offline and a second GAN technique. We called this dataset KDEF_Q.

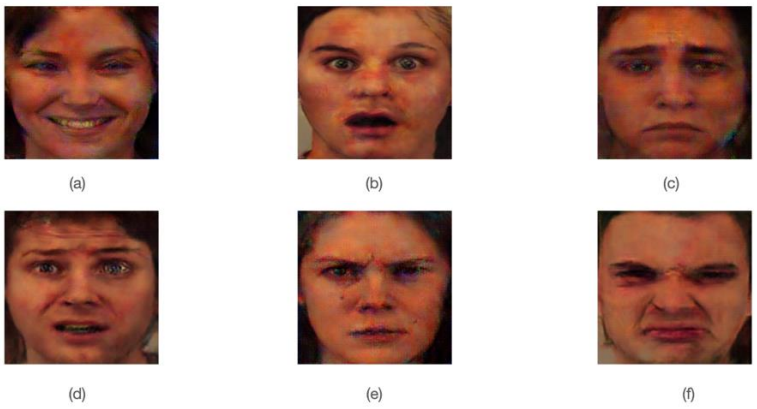


Figure 4. Examples of generated images by DCGAN models showing different face emotions: (a) happy, (b) surprised, (c) sad, (d) afraid, (e) angry, (f) disgusted

We add the following remarks. Since the activation function of the last layer of the generator applies the hyperbolic tangent, we rescale the input images between -1 and 1 before training the DCGAN model. Thus, we subsequently rescale once again the generated images between 0 and 255, in agreement with the adopted standardization. Furthermore, we note that another stable configuration for generator model takes a batch normalization layer after the first two transposed convolution layers and, at the same time, a global average pooling 2D layer instead of the flatten one at the end of the discriminator model. In both cases the quality of the produced images is quite sufficient. As shown in Figure 4, we get images of people whose features clearly express typical facial emotions. Although the quality of the generated images it is not comparable with the latest GAN techniques as [29][30], we test these images with an emotion classifier in order to check if the predicted emotions coincide with those of the images. We refer the discussion of this test to Section 5.3.

In addition to the three previous dataset we also consider their union, namely a dataset containing the original KDEF, the augmented offline version, the augmented GAN version obtained by [8] and our augmented GAN version. This dataset contains 7910 images and has been called KDEF_PFA_Q. Summarizing, the datasets on which we will train and test the emotion classifiers are listed in Table 1.

Table 1. Main features of train and test dataset

Dataset	Images	Usage
KDEF_OL	5880	Train
KDEF_PFA	6860	Train
KDEF_Q	6930	Train
KDEF_PFA_Q	7910	Train
CK+	902	Test
JAFFE	213	Test

4. MODELS AND TRAINING ALGORITHM

In this work we fine tune pre-trained models applying the transfer learning technique. We consider four deep learning architectures, each of which has placed a milestone in the problem of image recognition, namely the VGG16, VGG19 [21], InceptionV3 [31][22] and InceptionResNetV2 [32][23]. The models where trained on the ImageNet ILSVRC-2012 dataset (http://image-net. org/challenges/LSVRC/2012/), which includes more than one million images distributed along one thousand different classes [11].

The idea behind transfer learning consists in reusing the knowledge learned in solving a given problem and transferring it to solve a different but similar one [19] [11]. We consider deep networks already trained in a very big dataset to solve a problem of image classification. We thus reuse this knowledge, namely the values of the weights of the networks, to solve the problem of emotions classification. Furthermore, it is known that each layer of a neural network learns how to identify the features that are necessary to perform the final classification. Usually, lower layers identify lower-order features such as colors and edges, and higher layers compose these lower-order features into higher order ones such as shapes or objects. Hence, the intermediate layer has the capability to extract important features from an image which are useful for making a different kind of classification. Thus, to specialize the networks to our task, we applied fine-tuning technique freezing the values of the weights of a first part of the layers, and training the second part on ours specific dataset.

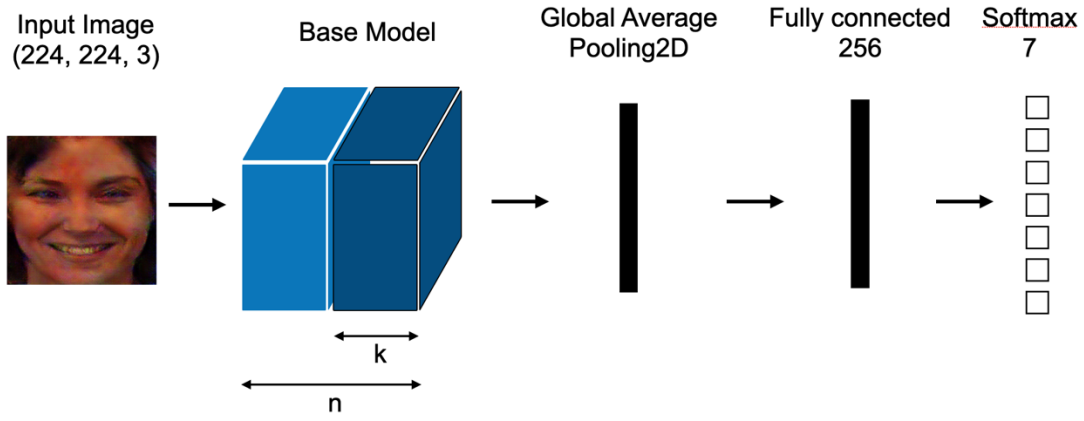


Figure 5. Architecture of neural networks used for transfer learning and fine tuning techniques. In the Base Model block, *n* represent the total number of layers, while *k* is the number of trainable layers during the fine tuning procedure.

4.1. Architecture of the models

We build the models joining the components one after the other. As shown in Figure 5, the first block is the base model, namely one of the pretrained networks. Tensorflow allows to download the base architecture where the weights are already trained in the ImageNet dataset. Since we want to fine tune the models in different datasets respect to the ImageNet one and with a different number of target classes, we specified the clause *include top=False* when downloading the model in order to remove the last layers related to the ImageNet classification task. Thus, we applied a global average pooling 2D layer, then a fully connected layer with 256 neurons and, lastly, a softmax activation function with seven outputs, corresponding to the seven possible emotions.

4.2. Fine tuning procedure

We divide the training procedure in two steps. Looking at the Figure 5, first we freeze the values of the weights of all the *n* layers of the base model, training for 10 epochs only the weights of the fully connected layer and the softmax function at the output. In this stage, the models are trained using the Adam optimizer with a learning rate equal to 10^{-3}. Subsequently, we unlock the last *k* layers of the base model, allowing to tune their weights on the training datasets for 65 epochs. In this stage, we always use the Adam optimizer but with a learning rate equal to 10^{-4} to not move too far from the optimal position.

As shown in Table 2, each base model has a different number of layers, thus a different number of parameters. To find the optimal value of tuned layers *k* in the second step, we carry out some train and validation test on the training datasets. Thus, for each of model, we choose the value of *k* maximizing the validation accuracy.

Table 2. Features of the base models

Base model	Layers	Tuned Layers	Params	Trainable params
VGG16	19	5	14714688	7079424
VGG19	22	9	20024384	14158848
InceptionV3	311	140	21802784	16215936
InceptionResNetV2	780	371	54336736	40442464

4.3. Images normalization

The different base models were pre-trained on the ImageNet dataset using different normalizations for the input images. We must therefore adapt our datasets to these normalizations. Thus, for the InceptionV3 and InceptionResNetV2 models, we scale the pixel intensity between 0 and 1. Instead, for the VGG16 and VGG19 models, first we convert the input images from RGB to BGR, then we zero-center each color channel with respect to the ImageNet mean, namely (103.939,116.779,123.68), without scaling.

5. EXPERIMENTAL SETUP AND RESULTS

We perform experiments following two strategies: a (i) *cross-datasets* approach, in which the model is trained and tested using different datasets, and (ii) *intra-datasets* approach, in which a global dataset is created by the union of the specific datasets and it is used for training and test.

5.1. Cross-datasets test

We implement the cross-datasets procedure by training the models on the datasets KDEF_OL, KDEF_PFA, KDEF_Q and KDEF_PFA_Q, then testing them on the CK+ and JAFFE datasets. To reduce the influence of random weights initialization, each architecture was trained and tested ten times. Thus, for each metric, we evaluate mean and standard deviation. The results for the accuracy values are summarized in Table 3.

Table 3. Mean accuracy and standard deviation of the 10 runs for each model on the CK+ and JAFFE databases

Train Dataset	Model	CK+ (%)	JAFFE (%)
KDEF_OL	VGG16	72.96 ± 8.55	42.72 ± 3.83
	VGG19	80.99 ± 6.99	39.12 ± 5.52
	InceptionV3	58.73 ± 8.07	42.66 ± 3.32
	InceptionResNetV2	81.28 ± 4.65	39.23 ± 4.15
KDEF_PFA	VGG16	68.44 ± 6.51	45.26 ± 3.53
	VGG19	66.34 ± 8.65	45.12 ± 4.12
	InceptionV3	50.60 ± 7.30	44.20 ± 2.56
	InceptionResNetV2	78.73 ± 6.71	45.21 ±3.09
KDEF_Q	VGG16	74.29 ± 5.59	40.61 ± 4.98
	VGG19	81.54 ± 4.47	37.15 ± 2.91
	InceptionV3	69.6 ± 8.37	43.19 ± 4.01
	InceptionResNetV2	**86.15 ± 3.54**	42.58 ± 3.86
KDEF_PFA_Q	VGG16	72.66 ± 5.40	46.10 ± 3.42
	VGG19	71.93 ± 5.31	42.91 ± 7.60
	InceptionV3	55.88 ± 5.74	**47.56 ± 2.41**
	InceptionResNetV2	79.76 ± 4.53	44.84 ± 4.11

The results vary significantly between the two test databases. The ability of the models to generalize on new images is quite high on the CK+, while it lowers considerably on the JAFFE. This fact shows the importance to test the models in at least two different databases to measure their generalization ability in the cross-database protocol. Furthermore, we observe that the generalization ability is also conditioned by the similarity between the test and train datasets. Actually, we cannot ignore that the JAFFE dataset is highly biased in term of gender and ethnicity, namely it comprises only Japanese female subjects.

The tests performed on the CK+ set show that the InceptionResNetV2 architecture not only achieves the highest accuracy on the KDEF_Q train set, with a mean value of 86.15% and a max peak of 90.35% between the ten runs, but it is also the most stable model because the smallest range of variation. On the other way, the InceptionV3 model fails to generalize. The results change slightly in the case of the JAFFE test set. In this case the InceptionV3 is the best model on each train datasets and reaches the maximum value for the accuracy when is trained on the dataset KDEF_PFA_Q.

Table 4. Comparison of the accuracy values between our InceptionResNetV2 model trained on the KDEF_Q database and other models tested on the CK+ database

Method	Training Dataset	Accuracy (%)
Proposed	Augmented KDEF	86.15
Porcu et al.[8]	Augmented KDEF	83.30
Zavarez et al. [7]	Six databases	88.58
Hasani et al.[33]	MMI + FERA	73.91
Lekdioui et al.[34]	KDEF	78.85

Table 5. Comparison of the accuracy values between our InceptionV3 model trained on the KDEF PFA Q database and other models tested on the JAFFE database

Method	Training Dataset	Accuracy (%)
Proposed	Augmented KDEF	47.56
Zavarez et al. [7]	Six databases	44.32
Ali et al.[33]	RaFD	48.67
Da Silva et al.[4]	CK	42.30

In Table 4 and Table 5, we compare the accuracy achieved by the proposed best architectures with those achieved by state of the art cross-database experiments conducted for FER systems and tested, respectively, on the CK+ and JAFFE database. In the case of the CK+ test set, our result is second only to the approach proposed by Zavarev et al.[7], that trained a VGG16 model in a dataset composed by six different database enlarged by using geometrical and colour transformation, therefore on a number of images higher than ours. On the other hand, we can see that our result slightly improves that obtained by Porcu et al.[8]. Actually, our approaches are quite similar. They differ in the way we applied the GAN techniques to increase the number of images of the KDEF dataset and, of course, the model architectures. Namely their analysis uses only a VGG16 model.

Table 6. Precision and recall values of the best InceptionResnetV2 model trained on the KDEF_Q dataset and applied to the CK+ dataset

Emotion	Precision (%)	Recall (%)
Angry	89	38
Disgust	72	98
Fear	55	48
Happy	100	93
Neutral	96	96
Sad	48	71
Surprise	92	93

Finally, we present the values of the metrics precision and recall computed for the single emotion classes when considering the best combination of model architecture. In Table 6 we show the results obtained for the InceptionResNetV2 model trained on the KDEF Q dataset and tested on the CK+ for which we get a peak of accuracy of 90.35%. We remember that the precision for a given class measures the number of correctly predicted samples out of all predicted samples in that class. Instead, the recall for a given class measures the number of correctly predicted samples out of the number of actual samples belonging to that class. Thus, although the InceptionResnetV2 applied to CK+ is a good classifier, it has a high false positives rate on fear and sad faces because roughly half of samples predicted as fear or sad actually do not belong to these classes. At the same time, it struggles to recognize angry and fear faces because only the 38% and 48%, respectively, of angry and fear faces are correctly classified. As happened for the accuracy, the precision and recall values of the best model evaluated on the JAFFE dataset get drastically worse. This is because the training datasets differ greatly from the JAFFE one, which containing only Japanese female subjects. Thus, the results obtained in this case are not of great relevance and it is useless to show them.

5.2. Intra-datasets test

The last experiment that we conduct is made on the union of all the datasets consider so far, namely, referring to Section 2, the original KDEF dataset, its augmented versions KDEF_DA_OL, KDEF_GAN_PFA, KDEF_GAN_Q and, finally, the CK+ and JAFFE databases. In this dataset, made of 9025 images, we apply a k-fold cross validation with k equal to 5. Using the same training procedure and the same model architectures of the previous section, we get the results showed in Table (7) for the accuracy on the validation folds.

Table 7. Mean accuracy and standard deviation on the five validation folds

Model	Accuracy (%)
VGG16	85.00 ± 1.83
VGG19	97.61 ± 0.58
InceptionV3	97.49 ± 1.83
InceptionResNetV2	97.99 ± 1.07

As we expected, the accuracy values obtained in the intra-database protocol are greater then those obtained in the extra-database protocol. Once again, the InceptionResNetV2 remains the model with the greatest accuracy with a mean of 97.99%. We also note that the VGG16 models is unable to reach the same accuracy values as the other models.

5.3. Quality test for GAN generated images

Table 8. Accuracy values for the best VGG19 and InceptionResnetV2 model trained on KDEF_OL dataset and tested on each separated group of 150 generated fake images

Emotion	VGG19 (%)	InceptionResnetV2 (%)
Angry	91	100
Disgust	87	88
Fear	97	85
Happy	95	97
Neutral	93	98
Sad	96	36
Surprise	97	99

In Section 3.2 we generated 150 fake images for each emotion by means of GAN techniques. In order to test their quality, namely to check if each group of images is actually classified as belonging to that class, we use the best emotion classifiers that we found in Section 5.1. In particular, looking to Table 7, we consider the models VGG19 and InceptionResNetV2 with the highest accuracy and trained in the dataset KDEF_OL because it does not contain the generated fake images that we want to test. The results are shown in Table 8. As learned in Section 5.1, the classifiers struggle to recognize sad faces, in fact only the 36% of them are correctly recognized by the InceptionResnetV2 model. In all other cases the generated fake images are correctly classified with accuracy values higher the 85%, thus their quality is quite good for artificially increasing the original KDEF dataset.

6. CONCLUSIONS

In this paper we extensively investigated various techniques in order to build efficient supervised systems able to recognize human face expressions. The main obstacle to solve this task is the

small size of the available training datasets. To get around this problem we made use of data augmentation techniques, such as geometrical transformations and training from scratch GAN models.

The experiments conducted in the cross-database protocol showed that the pretrained InceptionResnetV2 network, once fine tuned on an enlarged version of the KDEF database and tested on the CK+ test set, reaches a mean accuracy value of 86.15% with a close range of variation. Although the high values achieved for the accuracy, the model seem to suffer in recognizing emotions like fear and sad, for which we have obtained values of precision and recall under 70%. This problem is quite common to other FER systems and is probably related to the shape of faces that sometimes is very similar for these types of emotions.

The main obstacle to further increase the performances of the models and their ability to disentangle the recognition of face emotions remains the size and the composition of the training datasets. We showed that, even with few images for the training phase, GAN model can be built from scratch to obtain new synthetic images which are undoubtedly useful for obtaining the final performing FER systems. On the other hand, these synthetic images are of poor quality and all share the same appearance. In a future work, for the data augmentation step, we will apply transfer learning and fine tuning techniques to pretrained GAN models [36][37]. Thus, also with limited data, we will specialize GAN architectures to generate a large number of high quality images for each emotion type to be used, later, in the task of training an emotion recognizer.

REFERENCES

[1] Ekman, P., and Friesen, W. V. (1971). “Constants across cultures in the face and emotion”. Journal of Personality and Social Psychology, 17(2), 124-129.

[2] Canedo, Daniel and Neves, Antonio. (2019). “Facial Expression Recognition Using Computer Vision: A Systematic Review”. Applied Sciences. 9. 10.3390/app9214678.

[3] Li, Shan and Deng, Weihong. (2018). “Deep Facial Expression Recognition: A Survey”. IEEE Transactions on Affective Computing. PP. 10.1109/TAFFC.2020.2981446.

[4] Silva, Flavio and Pedrini, Helio. (2015). “Effects of cultural characteristics on building an emotion classifier through facial expression analysis”. Journal of Electronic Imaging. 24. 023015. 10.1117/1.JEI.24.2.023015.

[5] Caifeng Shan, Shaogang Gong, Peter W. McOwan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,”Image and Vision Computing, Volume 27, Issue 6,2009.

[6] Lopes, Andre and Aguiar, Edilson and De Souza, Alberto and Oliveira-Santos, Thiago. (2016). “Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order”. Pattern Recognition. 61. 10.1016/j.patcog.2016.07.026.

[7] M. V. Zavarez, R. F. Berriel and T. Oliveira-Santos, “Cross-Database Facial Expression Recognition Based on Fine-Tuned Deep Convolutional Network” 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Niteroi, Brazil, 2017, pp. 405-412, doi: 10.1109/SIBGRAPI.2017.60.

[8] Porcu, Simone and Floris, Alessandro and Atzori, Luigi. (2020). “Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems”. Electronics. 9. 10.3390/electronics9111892.

[9] P. Liu, S. Han, Z. Meng and Y. Tong, “Facial Expression Recognition via a Boosted Deep Belief Network,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1805-1812, doi: 10.1109/CVPR.2014.233.

[10] Yu, Zhiding and Zhang, Cha. (2015). “Image based Static Facial Expression Recognition with Multiple Deep Network Learning”. 435-442. 10.1145/2818346.2830595.

[11] Kapoor, Amita and Guili, Antonio and Pal, Sujit. (2019). “Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API,” 2nd Edition.

[12] M. Cakar, K. Yildiz, and Ö. Demir, "Creating Cover Photos (Thumbnail) for Movies and TV Series with Convolutional Neural Network," *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1-5, doi: 10.1109/ASYU50717.2020.9259872.

[13] K. Yildiz, E. Gunes, A. Bas (2021). CNN-based Gender Prediction in Uncontrolled Environments. Düzce Üniversitesi Bilim ve Teknoloji Dergisi. 890-898. 10.29130/dubited.763427.

[14] Goeleven, Ellen and De Raedt, Rudi and Leyman, Lemke and Verschuere, Bruno. (2008). "The Karolinska Directed Emotional Faces: A validation study, " COGNITION AND EMOTION.

22. 1094-1118. 10.1080/02699930701626582.

[15] Lucey, Patrick and Cohn, Jeffrey and Kanade, Takeo and Saragih, Jason and Ambadar, Zara and Matthews, Iain. (2010). "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010. 94 - 101. 10.1109/CVPRW.2010.5543262.

[16] F. Lin, R. Hong, W. Zhou and H. Li, "Facial Expression Recognition with Data Augmentation and Compact Feature Learning," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018, pp. 1957-1961, doi: 10.1109/ICIP.2018.8451039.

[17] Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Y..(2014). "Generative Adversarial Networks. Advances in Neural Information Processing Systems". 3. 10.1145/3422622.

[18] W. Yi, Y. Sun and S. He, "Data Augmentation Using Conditional GANs for Facial Emotion Recognition," 2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama), Toyama, Japan, 2018, pp. 710-714, doi: 10.23919/PIERS.2018.8598226.

[19] Yosinski, Jason and Clune, Jeff and Bengio, Y. and Lipson, Hod. (2014). "How transferable are features in deep neural networks?," Advances in Neural Information Processing Systems (NIPS). 27.

[20] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Li, Fei-Fei. "ImageNet: a Large-Scale Hierarchical Image Database, " IEEE Conference on Computer Vision and Pattern Recognition.

[21] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition, " International Conference on Learning Representations, 2015

[22] Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, ZB.(2016). "Rethinking the Inception Architecture for Computer Vision". 10.1109/CVPR.2016.308.

[23] Christian Szegedy and Sergey Ioffe and Vincent Vanhoucke and Alex A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence

[24] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding facial expressions with Gabor wavelets," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 200-205, doi: 10.1109/AFGR.1998.670949.

[25] OpenCV. (2015). Open Source Computer Vision Library.

[26] Wolf, Lior and Hassner, Tal and Maoz, Itay. (2011). Face recognition in unconstrained videos with matched background similarity. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 529 - 534. 10.1109/CVPR.2011.5995566.

[27] Radford, A., Metz, L., and Chintala, S. (2016)."Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". CoRR, abs/1511.06434.

[28] Kingma, Diederik and Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

[29] Karras, Tero and Laine, Samuli and Aila, Timo. (2019). "A Style-Based Generator Architecture for Generative Adversarial Networks". 4396-4405. 10.1109/CVPR.2019.00453.

[30] Karras, Tero and Laine, Samuli and Aittala, Miika and Hellsten, Janne and Lehtinen, Jaakko and Aila, Timo. (2019). "Analyzing and Improving the Image Quality of StyleGAN". https://arxiv.org/abs/1912.04958

[31] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[32] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[33] Hasani, Behzad and Mahoor, Mohammad. (2017). “Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields”. 790-795. 10.1109/FG.2017.99.

[34] Lekdioui, Khadija and Messoussi, Rochdi and Ruichek, Yassine and Chaabi, Youness and Touahni, Raja. (2017). “Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier.” Signal Processing: Image Communication. 58. 10.1016/j.image.2017.08.001.

[35] Ali, Ghulam and Iqbal, Muhammad and Choi, Tae-Sun. (2016). “Boosted NNE Collections for Multicultural Facial Expression Recognition. Pattern Recognition.” 55. 14-27. 10.1016/j.patcog.2016.01.032.

[36] Wang, Yaxing and Wu, Chenshen and Herranz, Luis and Weijer, Joost and Gonzalez-Garcia, Abel and Raducanu, Bogdan. (2018). “Transferring GANs: generating images from limited data.”https://arxiv.org/abs/1805.01677

[37] Karras, Tero and Laine, Samuli and Aittala, Miika and Hellsten, Janne and Lehtinen, Jaakko and Aila, Timo. (2020). “Training Generative Adversarial Networks with Limited Data.” https://arxiv.org/abs/2006.06676

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

PREDICTING ALZHEIMER'S DISEASE PROGRESSION BY COMBINING MULTIPLE MEASURES

Nour Zawawi[1], Heba Gamal Saber[2],
Mohamed Hashem[1] and Tarek F.Gharib[1]

[1]Department of Information Systems, Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt
[2]Geratic Medicine Department, Faculty of Medicine, Ain Shams University, Cairo, Egypt

ABSTRACT

Alzheimer's disease (AD) is a degenerative brain ailment that affects millions worldwide. It is the most common form of dementia. Patients with an early diagnosis of Alzheimer's disease have a strong chance of preventing additional brain damage by halting nerve cell death. At the same time, it begins to progress several years before any symptoms appear. The variety of data is the biggest problem encountered during diagnosis. Neurological examination, brain imaging, and often asked questions from his connected closed relatives are the three forms of data that a neurologist or geriatrics employs to diagnose patients. One of the biggest questions which need answering is the choice of a convenient feature.

The main objective of this paper is to help neurologists or geriatricians diagnose patient conditions. It proposes a new hybrid model for features extracted from medical data. It discusses AD's early diagnosis and progression for all features considered in the diagnosis and their complex interactions. It proves to have the best accuracy when compared with the state-of-the-art algorithm. Also, it proves to be more accurate against some recent research ideas. It got 95% in all cases, considering this work focused more on increasing the number of instances in comparison.

KEYWORDS

Alzheimer's disease, Diagnosis, Prediction, Classification, Feature Selection.

1. INTRODUCTION

Dementia is a general term used to describe symptoms that impact memory, the performance of daily activities, and communication abilities. Alzheimer's disease (AD) is the most common type of dementia. It gets worse with time and affects memory, language, and thought. As a result, it is a neurodegenerative disease described by progressive memory loss. It causes over 60% of dementia cases [1], [2]. Its patients usually have multiple symptoms. They range from a progressive loss of memory, language disorders, and disorientation. In general, there are several stages in AD. They are early, middle, and late (sometimes referred to as mild, moderate, and severe in a medical context) [3]. One of the main concerns facing specialists on the early detection of Mild Cognitive Impairment (MCI) is an intermediate stage between health and AD. It shows the potential of ongoing progression toward AD or other dementia. Although it does not

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 165-174, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111913

interfere with daily activities, it is abnormal given their age and education level. That is why it does not meet the criteria for AD.

Recent research shows that only 20–40% of individual cases will change to AD within three years; This is a lower rate of exchange reported in medical samples than in clinical cases [2]. However, AD's progression starts several years before any symptoms become visible and progressive [3]. Many drugs are in development, as there is no available treatment for AD [4], [5]. Researchers were able to diagnose Alzheimer's disease using modern diagnostic methods and biomarker tests. By combining biomarkers, it achieves varying levels of accuracy [4]. Unfortunately, the present research focuses on using MRI to classify illness states at their current stage rather than combining various features. As a result, these studies function as proof of concept without being tested in the real world [6], [7].

In the current age of artificial intelligence and machine learning technologies, predicting AD conversion is considered an important research area. The institutional use of machine learning techniques and the shift toward a personalized medicine concept, particularly in medical fields, represents a chance to improve therapeutic results. It makes personalized predictions with a high level of certainty based on the subject's specific data, which could help researchers and physicians make better and more effective judgments[8].

In this article, we propose a novel AD diagnosis method by combining multiple measures. Our measures include tests and MRI. For each measure, we extract all features that are shown as feature sets, respectively. Therefore, each one within ranked by accuracy in descending order. The top-ranked features of each feature set are applied to the multi-layer perceptron rule to obtain the best classification accuracy. After achieving the best accuracy, we can get the optimal feature subset. Afterward, to investigate the performance with chosen features. We propose a combined classifier to achieve AD classification. The rest of the paper is arranged as follows: Section 2 discusses the previous work related to our objective. The proposed model is illustrated in section 3. Section 4 shows the experiments made to achieve an objective. Finally, the discussion is shown in section 5.

2. LITERATURE REVIEW

Most studies on Alzheimer's disease (AD) have focused on using medical imaging as the only factor. Martı-Juan et al. [9] is a survey concentrating on longitudinal imaging data. It focused on papers that have published between 2007 and2019. Hong et al. [10] introduce Long short-term memory (LSTM) to predict AD development. It carries out the future state prediction for the disease rather than the state of a current diagnosis. While Janghel develops and compares different methods to diagnose and predict AD using MRI scans only [11]. It implements one model, which is the convolution neural network (CNN). At the same time, it uses four different architectures of CNN. An embedded feature selection method based on the least-squares loss function and within-class scatter for selecting the optimal feature subset is proposed by Cai et al. [12]. The optimal subsets of features used for binary classification are based on a support vector machine (SVM). Also, deep learning technology was discussed by Bi et al.[13]. It focused on the problem of automatic prediction of AD based on MRI images. It applies two main steps: 1- implement the unsupervised CNN for feature extraction. 2- utilizes the unsupervised predictor to achieve the final diagnosis.

According to our knowledge, Grassi et al. [14] use a weighted rank average grouped by different supervised machine learning methods to predict three-year conversion. Only a limited set of diverse characteristics are used to make predictions. The employment of algorithmic decision-making tools is the critical benefit. Liu et al. [15] provide a new method for detecting AD

based on spectrogram characteristics collected from voice data. It can assist families in better understanding the progression of a patient's sickness at an early stage.

Guo et al. [16] forecast the conversion from MCI to AD efficiently. Researchers proposed a cerebral similarity network containing more progression information. The classifiers were trained and evaluated using leave-one-out cross-validation and support vector machine (SVM). With a high accuracy of 92.31%, the proposed methodology was shown to be effective.

The following studies serve as the foundation for our study. In ascending order, they are listed. The first, [17], explains how MRI data can improve the accuracy of diagnoses for the Mini-Mental State Examination (MMSE) and logical memory (LM)tests. It accesses model correctness via Multilayer Preceptor. The second, [14], shows how clinically translatable strategies for conversion can be predicted. It also detects high-risk people who are converted. It continues to work three years after the initial assessment. Then, Haaksma et al. [18] address the link between Alzheimer's disease and its predictors. It included some Alzheimer's disease cases that have had at least one examination following diagnosis. To determine whether there are latent classes of Mini-Mental State Examination (MMSE) and Clinical Dementia Rating sum of boxes (CDRsb) routes across time. It employs growth mixture models with parallel processes. To find baseline predictors of class membership, researchers utilized bias-corrected multinomial logistic regression. A multimodal data [19] classifier that employs a hybrid deep neural network classifier. It is based on a set of MRI pictures as well as EEG inputs. The goal is to improve the learning process by incorporating the weight component of DNN into CNN. Then it explains how the accuracy of hybrid classifiers is determined.

To find correlations between brain areas and genes, use the appropriate correlation analysis approach at the conclusion. [20] was proposed via a cluster evolutionary random forest (CERF). It adds the concept of clustering evolution to increase the random forest's generalization performance. Farouk and Rady [21] in 2020 investigated the use of unsupervised clustering methods for the early identification of Alzheimer's disease. Though classification techniques are used to identify medical disorders, the lack or inaccuracy of labeled data might be an issue. This study compares the k-means and k-medoids using Voxel-Based Morphometry (VBM) characteristics taken from MRI scans.

3. METHODOLOGY

The feature selection technique is a knowledge discovery tool that helps grasp a problem by analyzing the most critical aspects. It seeks to improve classifiers by listing essential features, which also helps to reduce computational load. Due to the high correlation between features, many equally ideal signatures are commonly produced, making standard feature selection methods unstable and reducing the confidence in selected features [18], [22]. This section describes the two-tier feature selection. The feature ranking stage employs information entropy (IE) that uses a filtering approach. The stage aims at ranking subsets of features based on high information gain entropy in decreasing order. Therefore, this stage aims to extend additional features that contribute to the relationship between alerts with better discriminative ability than the initially ranked features. [23].

This paper describes a method (IE-MLP) for disease diagnosis and prediction based on feature selection utilizing optimality criteria. The information entropy Multilayer Perceptron (IE-MLP) [23], [24], [25] method uses optimality criterion for feature selection. Its major objective is to make the doctor's assessment easier. Figure 1 shows the block diagram; it has two basic steps:

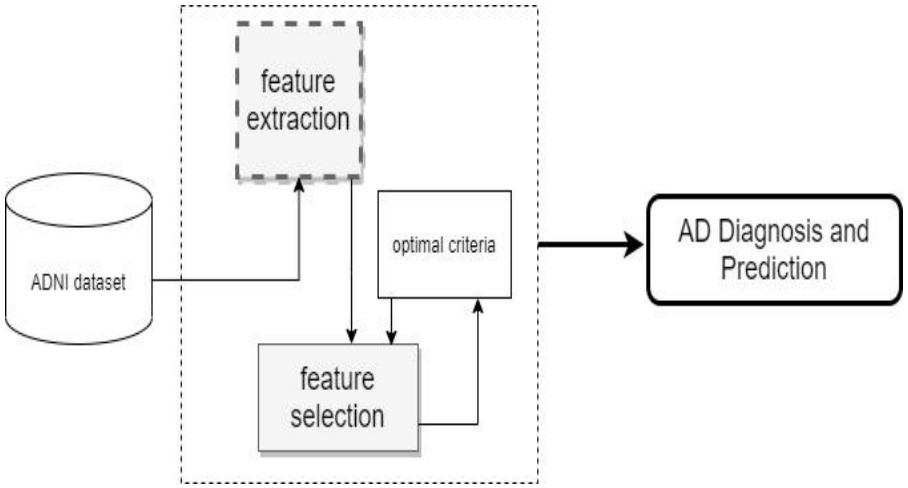


Figure 1: Block Digram of proposed Model

1) Feature extraction divides the data into three types in this phase (neurological test, MRI, diagnosis at baseline – the following section describes more details about the data). The doctor uses all three types of data at diagnosis. That is the reason behind selecting these various data sets kinds.

2) Feature selection reduces the number of features in a computation that uses many resources without losing important information. It makes the computer's task easier and faster; the importance of these steps is that it saves time and costs to increase diagnosis speed. It accepts the ADNI dataset's characteristics as input. Then, the Perceptron Learning Rule is applied to the smaller sets to determine the decision rule. It has a significant impact on disease diagnosis and prediction. The dataset goes through a pre-processing stage to alter data; the pre-processed data feds into the suggested model to extract relevant features. Perceptron rules are applied to the resultant feature vector to investigate the importance of various anatomical ROIs.

4. EXPERIMENTS

4.1. Dataset and Preprocessing

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database was utilized to compile the data for this article (adni.loni.usc.edu). The ADNI began as a public-private partnership in 2003. The primary purpose of ADNI is to identify clinical, imaging, genetic, and biochemical indicators for Alzheimer's disease early identification and tracking (AD). It includes many cognitively normal, MCI, and AD patients recruited from over 50 different US and Canadian facilities, with six-month follow-up examinations. The proposed work uses ADNIMERGE (this subset is part of the official dataset released by ADNI). It includes Clinical and biomarker data from the Alzheimer's Disease Neuroimaging Initiative. It contains 90 attributes and 12612 instances. Although it has nine classes, we include only three in this paper (AD, MCI, and NL).

As previously stated, the ADNI collection contains a variety of data kinds. By setting the duration for test retakes, this effort intends to aid diagnosis utilizing other data. For these reasons, the categories of data used are as follows:

- Neurological Examination: For specialized examination of the brain and mental health issues (neuropsychologist). Tests included assessing memory and cognitive abilities in the evaluation.
- The patients' initial tests and diagnoses serve as a baseline.
- Image processing in the brain: Only two categories of technologies are available in the ADNI dataset (MRI used only)

4.2. Feature Selection

Table 1: Feature Selection Comparison

Original Datset	Attribute Number
Baseline (19 attributes)	1
MRI (7 attributes)	1
Neurological (9 attributes)	7

As mentioned in the previous subsection, data contains 87 attributes and 2700 instances. After data pre-processing, it includes 45 attributes and 2700 instances. In the beginning, 45 attributes are a significant number to analyze. So, the feature extraction and selection model chooses more convenient features. At the same time, the chosen features need to be acceptable by medication standards.

The IE-MLP works as follows: 1) Information entropies arranges the data in ascending order depending on their values 2) Multilayer perceptron got the arranged data, and the suitable feature is selected. Table 1 shows the number of attributes before and after the proposed model. The difference between each model and accuracy is discussed in more detail in section 5.

4.3. Classification Technique

A neural network (NN) is a type of machine learning which models itself after the human brain; While the basic unit of the brain is a neuron, the essential building block of NN is a perceptron connected to an extensive network. It can perform deep learning. An ANN and error function helps calculate the gradient of a loss function for all the weights in the network. In this paper, a multi-layer perceptron (MLP) chose for the training model. The following are some of the reasons for choosing this model:

1. It is fast, simple, and easy to program
2. It is flexible as it does not require prior knowledge
3. It does not need any special mention of the features of the function to be learned.

The NN uses no hidden layer, and it contains seven input and three output units. With a Learning rate equal to 0.3. The activation function used is Approximate Sigmoid, and the loss function is squared error.

5. DISCUSSION

A confusion matrix describes the performance of a classification model (or "classifier"). Equations1:4 contains a set of rates lists that computes from a confusion matrix. Lets now define the most basic terms, which are whole numbers (not rates):

- True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- True negatives (TN): We predicted no, and they do not have the disease.
- False positives (FP): We predicted yes, but they do not have the disease.
- False negatives (FN): We predicted no, but they do have the disease.

Table 2: Confusion Matrix for 800 attributes

Measure	Value (%)
Accuracy	95.8
Error Rate	4.5
Specifity	95.5
Sensitivity	95.6
Roc-Curve	99

Table 3: Varying data size vs. Time

Data Size	Time (Sec)
800	2.28
700	2.04
600	1.71
500	1.43

The following criteria are the chosen ones from the confusion matrix. The reason behind selecting them is that they will have real meaning to our objective. According to equation1, accuracy is one metric for evaluating classification models. It is the ratio of the number of correct predictions to the total number of input samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} \qquad (2)$$

Equation 2 illustrates the Error Rate, which is the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0. As the error rate increases, the model reliability decreases.

Specificity (SP) is the number of correct pessimistic predictions divided by the total number of negatives. It showed in equation 3. It is also called actual negative rate (TNR)

$$Specifity = \frac{TN}{N} \qquad (3)$$

Precision discussed in equation 4; tries to answer what proportion of identifications was correct. Our model got 0,95. It means a high number of instances that were correctly classified.

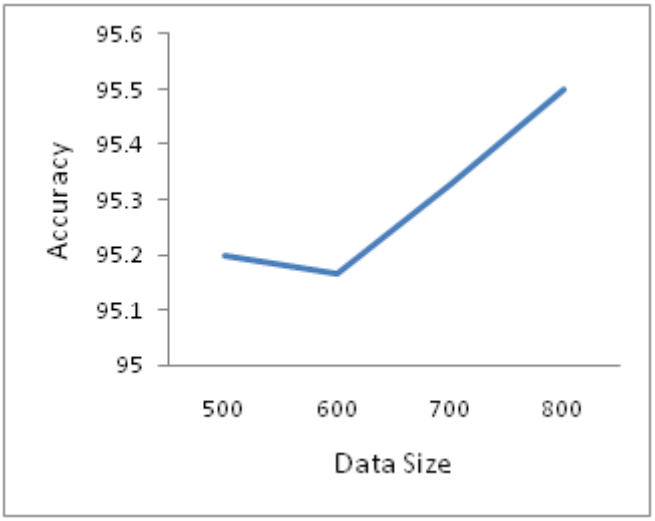


Figure 2. Varying Datasize Vs. Time

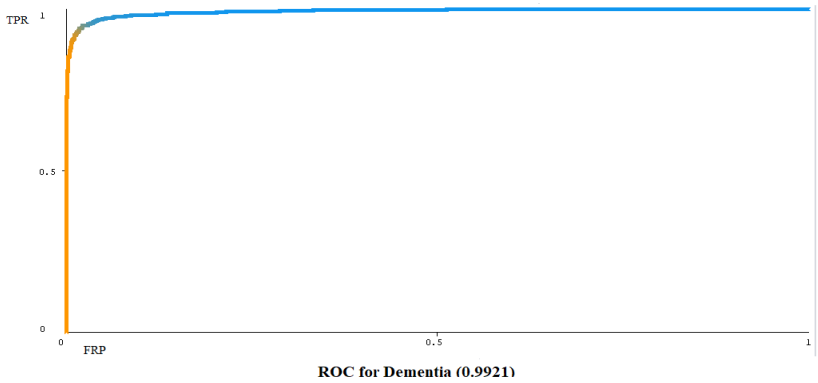


Figure 3. Area Under Curve-AUC for dementia

$$Sensitivity = \frac{TP}{TP + FN} \qquad (4)$$

ROC curves are a valuable tool for evaluating a diagnostic test's performance over a wide range of possible values for a predictor variable. The area under a ROC curve is a measure of discrimination that researchers can use to compare the results of two or more diagnostic tests. A ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. The ROC curve is a graph with:

- The x-axis shows 1 – specificity
- The y-axis shows sensitivity

In Order to view the difference based on varying data sizes, including data sizes from 500 to 800, the confusion matrix (equation 1:4) comparison with different feature types is taken in Table 2. The results show that our algorithm achieves the best performance with a chosen feature. Table 3 and figure 2 show that our model is stable in most of the features.

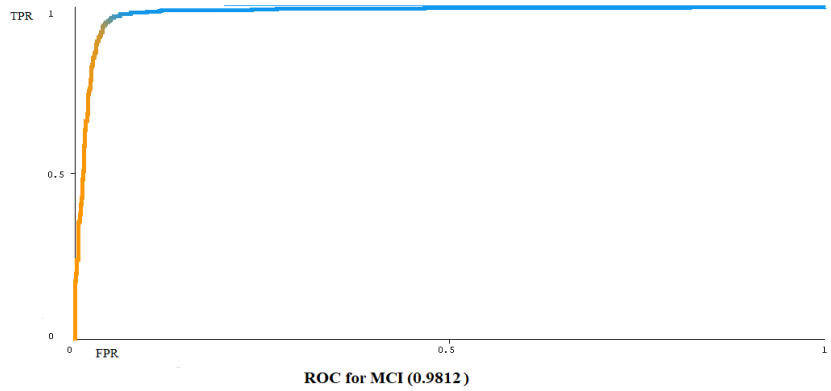


Figure 4. Area Under Curve-AUC for MCI

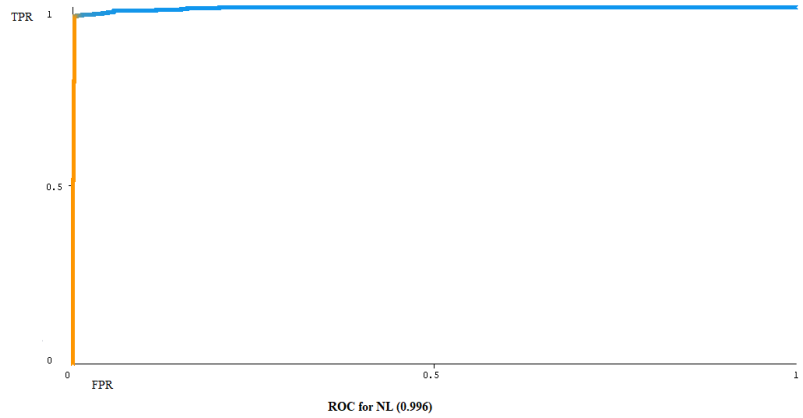


Figure 5. Area Under Curve-AUC for NL

ROC curves depict the relationship/trade-off between clinical sensitivity and specificity for each conceivable cut-off for a test or a set of tests in a graphical format. Furthermore, the area under the ROC curve provides insight into the value of using the test(s) in question. Figure 3, 4,5 shows the ROC area for three classes (Dementia, MCI, NL). The proposed model got a 0.99 average value for 800 class attributes. To indicate better performance by curves that are closer to the top-left corner [26]. It means that the model has a good measure of separability, which means there is a 99% chance that the model will distinguish between different cases. The next step in the proposed work is to apply the results in actual experiments to evaluate the model performance.

6. CONCLUSIONS

This study introduced a proposed framework for Alzheimer's disease diagnosis by combining multiple measures. It explores the effect of features extraction from MRI and neurological tests. Diagnosis of AD depends on multiple features that facilitate the process. The main objective of this study is to improve time and save time on specialties' needs. Experimental results on the ADNI database have shown that our proposed method is efficient and can achieve better classification performance. The proposed approach managed successfully to obtain an early AD diagnosis with an accuracy of 97%. The clustering techniques used in the proposed approach provided an automated preliminary insight discovering vital early patterns in the data with reliable accuracy. Our results further demonstrate that clinical AD diagnosis could benefit from calculating multiple measures from diagnostic data and incorporating these all in automated analysis.

REFERENCES

[1] J. Zeisel, K. Bennett, R. Fleminget al., "World Alzheimer report 2020: Design, dignity, dementia: Dementia-related design and the built environment," 2020.

[2] Association, "2019 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol. 15, no. 3, pp. 321–387, 2019.

[3] Patterson, "World Alzheimer report 2018: the state of the art of dementia research: new frontiers," Alzheimer's Disease International (ADI), pp. 32–36,2018.

[4] N. Davda and R. Corkill, "Biomarkers in the diagnosis and prognosis of Alzheimer's disease," Journal of Neurology, vol. 267, pp. 2475–2477, 2020.

[5] M. Buegler, R. L. Harms, M. Balasa, I. B. Meier, T. Exarchos, L. Rai, R. Boyle, A. Tort, M. Kozori, E. Lazarou, M. Rampini, C. Cavaliere, P. Vlamos,M. Tsolaki, C. Babiloni, A. Soricelli, G. Frisoni, R. Sanchez-Valle, R. Whelan, E. Merlo-Pich, and I. Tarnanas, "Digital biomarker-based individualized prognosis for people at risk of dementia," Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 12, no. 1, p. e12073, 2020. VOL. 14, NO. 8, AUGUST 20153

[6] R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, and W. Chan, "Predicting progression of alzheimer's disease," Alzheimer's Research &Therapy, vol. 2, no. 14, 2010.

[7] G.-R. J, ́Avila Villanueva M, and F.-B. M ́A, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," Sci Rep., vol. 26, no. 10, p. 20630, Nov. 2020.

[8] W. Y. Ng, C. Y. Cheung, D. Milea, and D. S. W. Ting, "Artificial intelligence and machine learning for alzheimer's disease: let's not forget about theretina," British Journal of Ophthalmology, 2021.

[9] G. Mart ́ı-Juan, G. Sanroma-Guell, and G. Piellaa, "A survey on machine and statistical learning for longitudinal analysis of neuroimaging data inalzheimer's disease," Computer Methods and Programs in Biomedicine, 2020.

[10] X. Hong, R. Lin, C. Yang, N. Zeng, C. Cai, J. Gou, and J. Yang, "Predicting alzheimer's disease using lstm," IEEE Access, vol. 7, pp. 80 893–80 901,2019.

[11] R. R. Janghel,Deep-Learning-Based Classification and Diagnosis of Alzheimer's Disease. IGI Global, 2020, ch. 76, p. 25.

[12] J. Cai, L. Hu, Z. Liu, K. Zhou, and H. Zhang, "An embedded feature selection and multi-class classification method for detection of the progression from mild cognitive impairment to alzheimer's disease," Journal of Medical Imaging and Health Informatics, vol. 10, no. 2, pp. 370–379, 2020.

[13] X. Bi, S. Li, B. Xiao, Y. Li, G. Wang, and X. Ma, "Computer aided alzheimer's disease diagnosis by an unsupervised deep learning technology," Neurocomputing, 2019.

[14] G. M, R. N, C. D, L. D, S. K, P. G, D. M, and A. D. N. Initiative, "A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures," FrontNeurology, vol. 10, no. 756, 2019.

[15] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying alzheimer's disease," Simulation Modelling Practice andTheory, vol. 99, 2020.

[16] M. Guo, Y. Li, W. Zheng, K. Huang, L. Zhou, X. Hu, Z. Yao, and B. Hu, "A novel conversion prediction method of mci to ad based on longitudinaldynamic morphological features using adni structural mris," Journal of Neurology, vol. 267, pp. 2983–2997, 2020.

[17] S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama, "Fusion of deep learning models of mri scans, mini-mental stateexamination, and logical memory test enhances diagnosis of mild cognitive impairment," Alzheimers Dement (Amst)., vol. 28, no. 10, pp. 737–749, Sep.2018.

[18] H. ML, Calder ́on-Larra ̃naga, O. R. MG, M. RJ, and L. JS, "Cognitive and functional progression in alzheimer disease: A prediction model of latentclasses," International journal of geriatric psychiatry, vol. 33, no. 8, 2018.

[19] A. Shikalgar and S. Sonavane, "Hybrid deep learning approach for classifying alzheimer disease based on multimodal data," in Computing in Engineering and Technology, B. Iyer, P. S. Deshpande, S. C. Sharma, and U. Shiurkar, Eds. Singapore: Springer Singapore, 2020, pp. 511–520.

[20] X. Bi, X. Hu, H. Wu, and Y. Wang, "Multimodal data analysis of alzheimer's disease based on clustering evolutionary random forest," IEEE Journal of Biomedical and Health Informatics, 2020.

[21] Y. Farouk and S. Rady, "Early diagnosis of alzheimer's disease using unsupervised clustering," International Journal of Intelligent Computing and Information Sciences, vol. 20, no. 2, pp. 112–124, 2020.
[22] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," Journal of King Saud University – Computer and Information Sciences, 2019.
[23] S. Nembrini, I. R. K¨onig, and M. N. Wright, "The revival of the gini importance?" Bioinformatics, vol. 34, no. 21, pp. 3711–3718, 2018.
[24] H. Sun, A. Wang, Q. Ai, and Y. Wang, "A new-style random forest diagnosis model for alzheimer's disease," Journal of Medical Imaging and Health Informatics, vol. 10, no. 3, pp. 705–709, 2020.
[25] Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," Journal of Neural Engineering, vol. 16, no. 3, p. 031001, apr 2019. [Online]. Available: https://doi.org/10.1088/1741-2552/ab0ab5
[26] Mandrekar, J. N. "Receiver Operating Characteristic Curve in Diagnostic Test Assessment" Journal of Thoracic Oncology, 2010, 5 , 1315-1316

AUTHORS

Nour Zawawi
PHD student , Department of Information Systems ,Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

Heba Gamal Saber,
Consultant & Lecturer of Geriatric Medicine - Ain Shams University
Geriatric Doctor Specialized in Elder Memory

Mohamed Hashem
Previous vice dean, Former Head of Information Systems Department Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

Tarek F.Gharib
Head of Information Systems Department
Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

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

INTERACTIVE DASHBOARD DESIGN FOR MANAGER, DATA ANALYST AND DATA SCIENTIST PERSPECTIVE

Temitope Olubunmi Awodiji

Department of Computer Information Science,
California Miramar University, California, USA

ABSTRACT

With large amounts of unstructured data being produced every day, organizations are trying to extract as much relevant information as possible. This massive quantity of data is collected from a variety of sources, and data analysts and data scientists use it to create a dashboard that provides a complete picture of the organization's performance. Dashboards are business intelligence (BI) reporting tools that collect and show key metrics and key performance indicators (KPIs) on a single screen, enabling users to monitor and analyse business performance at a glance. An objective assessment of the company's overall performance, as well as of each department, is provided. If each department has access to the dashboard, it may serve as a springboard for future discussion and good decision-making. The goal of this article is to explain in detail the implementation of Dashboard and how it works, which will serve as a blueprint for building an effective dashboard with respect to best practices for dashboard design.

KEYWORDS

ETL, Data, Dashboard, Data Analyst, data Science.

1. INTRODUCTION

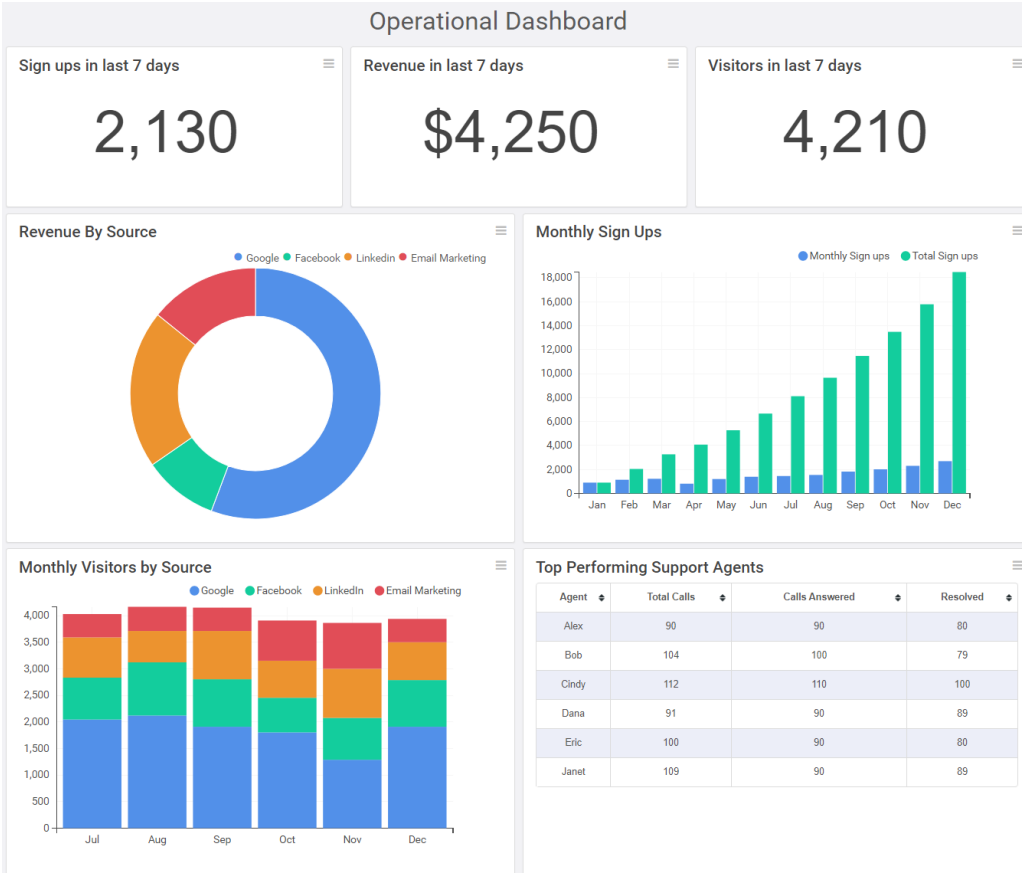
Definition - What does dashboard mean?

A Dashboard is said to be a simple to read, frequently single page, real-time user interface, demonstrating a graphical introduction of the present status (snapshot) and historical trends of an organizations or computer appliances key performance indicators to enable immediate and informed choices to be made briefly. There are a lot of ideas about what a dashboard is, which this article will clearly define.

There are typically four types of presentation media: dashboards, visual analysis tools, scorecards, and reports but our focus will be on dashboard designing. These are all visual representations of data that help people identify correlations, trends, outliers (anomalies), patterns, and business conditions. Similarly, they all have their own unique attributes. A dashboard in a simple word can provides us with a Summary of the status of a Performance of a project, Sales report, Customer details Updates, Employees report etc. A dashboard gives the real-time changes happening that can be analyse further. There are three most used dashboards among many other.

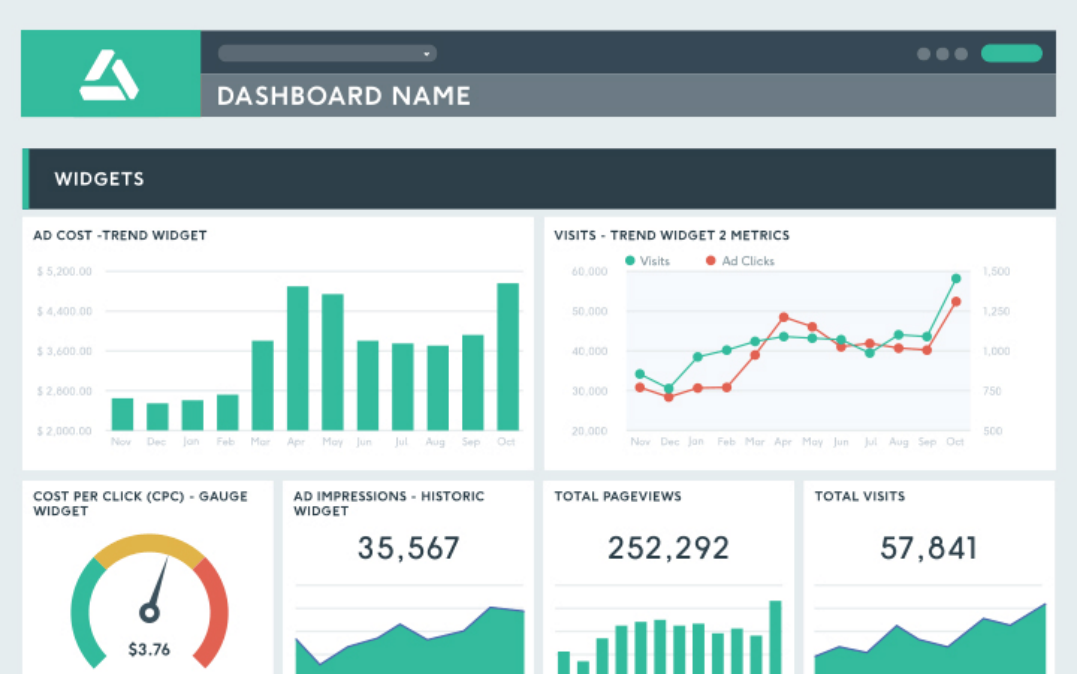
David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 175-185, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111914

The operational dashboard: This is said to be the simplest and most popular kind of dashboard that shows real-time changes in data for various operations. This is a visualization tool that helps to track business operations and provides us with an updated performance report. These dashboards are good at summarizing large amounts of data and are designed in a way such that they can be viewed multiple times a day, say to check their progress towards a target set. It is mostly used by managers to monitor the performance of their employees. These simple dashboards can collect data from a single or multiple sources and present it in a simple and readable at-a-glance format.



Source from: https://ubiq.co/analytics-blog/create-operational-dashboard-business/

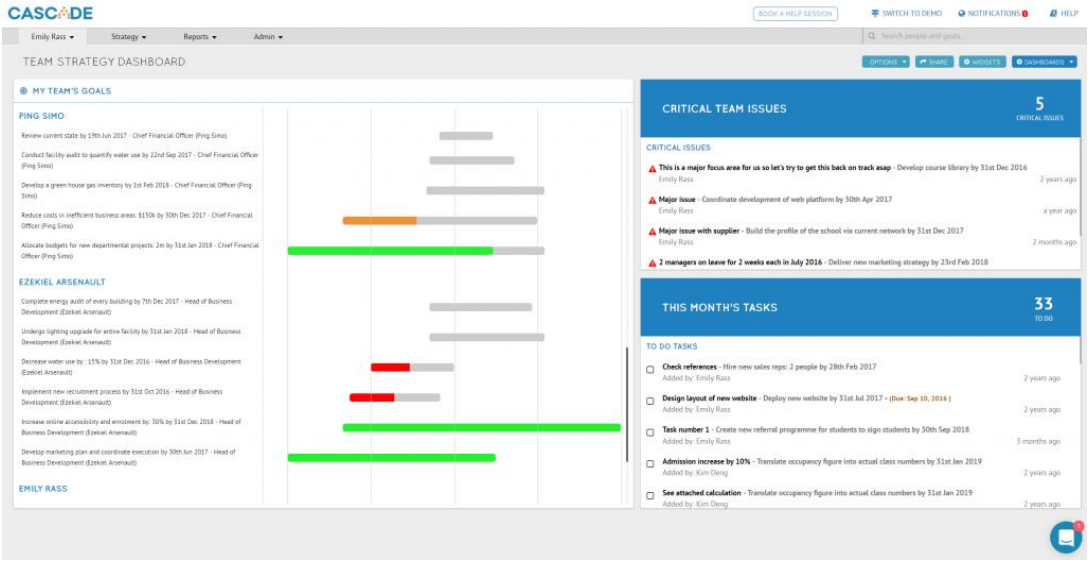
Analytical Dashboards: This is the types of dashboards that used data from previous reports to make further decisions. This is used to track changes, predict the result and conclusion as well. Drill through, Pivot Tables are some features that is available in this dashboard. This Dashboards is commonly used by Data Analyst who perform slicing of Data to compare and Investigate Occurring Outcomes.



Source from https://dashthis.com/analytical-dashboard/

Strategic Dashboards: These dashboards help the managers and stakeholders to determine the current state of the organization. With this dashboard, the area of Improvement can be detected and ways to avoid

business disruption can be established.



Source from: https://www.cascade.app/blog/examples-to-create-strategy-dashboards

Advantages of A Dashboards: A dashboard makes it easier to grasp data from different sources and view them at a single glance. It gives a clear picture of the company's progress and improves decision making. Dashboards are very important in today's industry. It allows the stakeholders as

well as the user to select certain key performance indicators to display and would provide real-time updates whenever needed. These indicators can be based on anything such as output, input, time, and activity. A good dashboard will provide key benefits such as total business visibility, significant time savings, reduced stress, increased productivity, and increased profits. The most important part of a good dashboard is the part that gets the least amount of attention, which usually shows the underlying data. There are several types of ways to set up dashboards, but they are all categorized into three, which will be explained further in the paper.

2. IMPLEMENTATION OF A DASHBOARD DESIGN

At a high level, it may seem relatively easy to build a dashboard. Companies that feel they have a good handle on which performance indicators are of strategic importance to the organization may think collecting, summarizing, and consolidating the supporting data shouldn't be that difficult. However, such oversimplification can lead to a failed project before it ever gets off the ground. A dashboard may appear to be a simple task at first glance. Organizations that believe they have a strong hold on which performance indicators are critical to their business operations may believe that gathering, summarizing, and aggregating the supporting data shouldn't be too tough. However, oversimplification can lead to a failed project. Successfully implementing a dashboard requires a step-by-step process and a methodology that considers all aspects of the project life cycle. This series of tasks includes planning, design, building, and deploying. This will be similar, regardless of the technology or vendor chosen. It is important to include all these steps above correctly to design and implement a dashboard that will have the potential to bring an immediate and considerable return on investment (ROI) to your organization.

2.1. Characteristics of a good dashboard design

A dashboard is an information management tool that visually tracks, explains, or break down and shows key performance indicators (KPI), metrics and key data points to monitor the wellbeing of a business, department, or explicit process. They are customizable to meet the specific needs of a department and organization.

- All the visualizations fit on a single PC screen — It should fit on one screen, but there may be scroll bars for tables with a lot of rows or charts with too many Data points.
- It demonstrates the most important performance indicators / performance measures to be checked.

- It should be able to discover correlations, trends, exceptions (irregularities), and business conditions in Data.
- Interactivity for example filtering and drill-down can be used in a Dashboard; however, those kinds of activities ought not be required to see which performance indicators are failing to meet expectations.
- It is not designed exclusively for decision makers but instead ought to be utilized by the general workforce as effective dashboards are straightforward and use.
- The displayed Data automatically refreshed with no help from the client. frequency of the update will differ by organization and by purpose. The most effective dashboards have Data updated on a minimum daily.

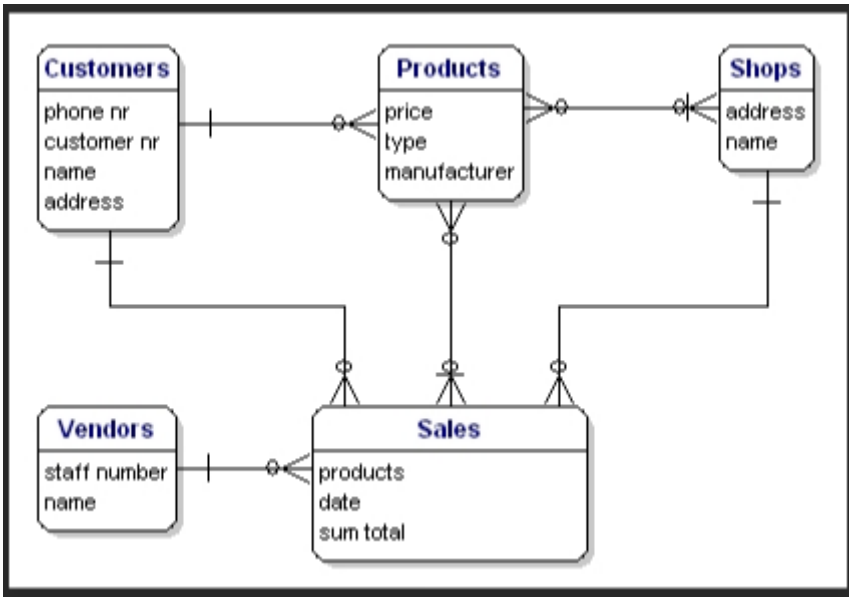
2.2. Database Vs Dashboard

A Database is a management framework for your data. Databases come in about as many various forms as the data comes in. A dashboard is a visual display of the most important of the most imperative data expected to accomplish at least one goals; combined and organized on a single screen so the information can be monitored briefly. We can say both has to do with information or data of your company.

Types of Databases

- Relational database.
- Flat-file database.
- Hierarchical & Network database.
- Object-oriented database.
- Object-relational database.

Relationships between Database and Dashboard



THE ABOVE DIAGRAM EXPLAIN HOW A DATABASE LOOK LIKE

THE ABOVE DIAGRAM EXPLAIN HOW A DASHBOARD LOOK LIKE

2.3. Identify the necessary internal and likely external databases that would be needed to support the Dashboard

After external database servers are defined, the databases on them can be defined. Altus Director can utilize databases that as of now exist on those servers, or it can create them while bootstrapping new Cloudera Manager examples or CDH clusters.

After outer database servers are characterized, the databases on them can be characterized. Altus. The following parts of an existing database must be defined which are additionally expected to support a dashboard:

- Type - The type of database, “MYSQL” or “POSTGRESQL.”
- Hostname – Server host name.
- Port – Server listening port.
- Name – The server database name.
- Username – User account name that has full access to the database.
- Password – User account password.

The parts of an external database template are:

- Name - A unique name for the template within the deployment or cluster template.
- Database Server Name - The name of the external database server where the new database is to reside.
- Database Name Prefix - The string prefix for the name of the new database server.
- Username Prefix - The string prefix for the name of the new user account that will have full access to the database.

The database server name in a database server template must refer to an external database server that is already defined.

At the point when Altus Director makes the new database, it names the database by beginning with the prefix in the template and after which appends a random string.

2.4. Internal database

Every organization has an internal network that sends and gets data from different sources inside and outside the organization. That network additionally stores all the internal information identified with organization like sales data, consumer feedback which are part of the data expected to help a database etc. Any such collection of data with respect information on market and consumer behaviour in electronic form will be called as internal database.

I. CRITICAL ENTITY RELATIONSHIPS (ER) ACROSS DATABASES TO ENSURE HIGH LEVEL OF INFORMATION ACCURACY

ER diagrams help ensure that the relationships among the data entities in a database are rightly organized so that any application programs developed are consistent with business operations and user needs. In addition, ER diagrams can also help as reference documents after a database is in use. To the database if changes are made, ER diagrams help design them.

A good example is an ER that is designed for an order database. In this database design, A salesperson will serve many customers. This is an example of a one-to-many relationship, as shown by planned data redundancy is a method of structuring data in which the logical database design is altered so that data entities are joined together, sum totals are taken in the data records instead of calculating from elemental data, and some data attributes are repeated in more than one data entity to for database performance improvement. data model - A diagram of data entities and their relationships. Enterprise data modelling is the Data modelling done at the level of all the enterprise. entity-relationship (ER)diagrams- Data models that use basic graphical symbols to show the organization relationships between data.

II. DESCRIBE THE DBMS/ERP RELATIONSHIPS IN HIGH-LEVEL TERMS THAT SUPPORT THE EXTRACT, TRANSFORM, AND LOAD (ETL) FUNCTIONS NECESSARY TO DELIVER INFORMATION AND FORMATS (E.G., CHARTS, ETC.) TO THE DASHBOARD

In computing, extract, transform, load (ETL) is a procedure in database utilization to prepare data for analysis, particularly in data warehousing which ended up well known during the 1970s.Data extraction clarifies extracting data from homogeneous or heterogeneous sources, while data transformation processes data by changing them into a legitimate storage design/structure for the questioning and analysis purposes.

Lastly, data loading explains the insertion of data into the final target database example is an operational data store, a data mart, or a data warehouse. An appropriately designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that different sources can be utilized together, and finally delivers data in a presentation-ready format so that application developers can design applications and end users can make decide.

2.5. Dashboard perspectives

I. KEY PRINCIPLES TO DESIGN A DASHBOARD

Building an effective dashboard with respect to best practices for dashboard design is the summit of a complete BI process would more often include gathering requirements, defining KPIs and creating a data model.

II. BUSINESS VALUE OBTAINED

There are five key benefits of Business Value that can be obtained from a Dashboards.

BVDs bridge the information gap.

BVDs channel various data points from within the business into a centralized location – which means geographical, organizational, production, IT and other data can be combined to gain a more contextual understanding of the information at your disposal which offers clear, colorful graphical user interfaces (GUI) that help users easily understand and evaluate complex datasets. Example: a car's dashboard.

A. Business value give business meaning to IT data.

It simplifies IT data into a format that is understandable to C-level staff members.

B. Visibility

BVD provides the organization with unparalleled visibility and insight and puts all the information at their fingertips.

C. With BVD you can gauge performance against your plan
D. They afford time savings for executives.

3. CRITICAL SUCCESS FACTORS

a. Your dashboard should provide the relevant information in about 5 seconds.

The excess of a dashboard designed should be able to answer most often asked business questions briefly.

b. Logical Layout: The Inverted Pyramid

Display the most significant insights on the top part of the dashboard, trends in the middle, and granular details in the bottom.

c. Minimalism: Less is More

Each dashboard should contain no more than 5-9 visualizations.

d. Choosing the right data visualization

Select the appropriate type of data visualization according to its purpose.

Before choosing a visualization, consider which type of information you are trying to relay:

- Relationship
- Comparison
- Composition
- Distribution

4. COMMON PITFALLS

Common pitfalls in dashboard design: -

I. A very common mistake is starting off with too much complexity. I ascribe to the KISS principle – Keep It very Simple and uncomplicated.

II. Taking it further, too much complexity can also lead to the data being separated into multiple screens or into different instances of a single screen.

III. Another common dashboard design blunder is to use gadgets or widgets like speedometers and gauges that may not lend any context to the data in terms of letting you see if your KPIs are on track, better than in the past or worse than projected.

IV. You can avoid using the same type of visual representation of a KPI multiple times for the sake of offering up more “visual variety” to viewers. At the bottom line, they should be able to assess different KPIs using a visual display that’s both intuitive and interactive for them.

I. Another common faux pas in designing visual representations of KPIs is not displaying data from the "zero" level, which can create a distorted view of the data and a significant discrepancy between the KPI's real and perceived values.

5. CONCLUSIONS

Choosing the correct visualization is key to making sure your end users understand what they are looking at, yet that is not all you ought to consider. When contemplating about how to design a dashboard you need to put in consideration who will be the end user of the dashboard in any case. For instance, when designing a dashboard for an end user focused on ad platform optimization, you might need to focus on metrics that will increase conversion rates. Because your end user will be in the loop of what goes on with every ad on a day-to-day level, considering the nitty-gritty measures such as CPM (cost per Mille) would make more sense and choosing an ERP system is a very difficult thing to experience. It required the knowledge and understanding of a variety of business units. On the other side, choosing DBMS is as good as the experience you go through with ERP system chosen. But an organization should not choose an ERP without considering possible underlying DBMS's. DBMS vendors always guaranteed to be the best around and they all can also give out a list of notes of their satisfied customers. Implementing ERP is a key decision, which involves human as well as monetary resources and requires legitimate assessment and business process re-designing. If you are unable to track your progress at work whenever your organization needs Improvement, and you are also unable to determine in which area, if your Data are stored in multiple places and you are finding it difficult to compare or analyze them, then you should consider dashboard Creation which will make your Data Analysis and Management a lot easier and your life better.

ACKNOWLEDGEMENTS

I want to express my profound gratitude to AIAA for the opportunity given to me for Publishing My Paper in this Journal.Many thanks to Richard Arogundade for his Tremendous help towards the Success of this Journal.

REFERENCES

[1] Bachman, Charles W. (1973). "The Programmer as Navigator". Communications of the ACM. 16 (11): 653–658. doi:10.1145/355611.362534.

[2] Beynon-Davies, Paul (2003). Database Systems (3rd ed.). Palgrave Macmillan. ISBN 978-1403916013.

[3] Chapple, Mike (2005). "SQL Fundamentals". Databases. About.com. Archived from the original on 22 February 2009. Retrieved 28 January 2009.

[4] Retrieved from https://www.cloudera.com/documentation/director/latest/topics/director_external_db_using.html

[5] Retrieved from https://www.silvon.com/blog/dashboard-design-5-pitfalls-avoid/ Denney, MJ (2016). "Validating the extract, transform, load process used to populate a large clinical article database". International Journal of Medical Informatics. 94: 271. doi:10.1016/j.ijmedinf.2016.07.009. PMC 5556907. PMID 27506144.

[6] Retrieved from https://www.pluscharts.com/dashboard-need-and-purpose/

[7] Retrieved From https://www.google.com/search?q=sample+Strategic+dashboards+images&tbm=isch&ved=2ahUKEwiv2uPIh5n0AhUMQkIHHbb1D2AQ2-cCegQIABAA&oq=sample+Strategic+dashboards+images&gs_lcp=CgNpbWcQA1CLDViCT2DaU2gAcAB4AYABygaIAeUhkgEOMTAuNi4xLjEuMC4yLjGYAQCgAQGqAQtnd3Mtd2l6LWltZ8ABAQ&sclient=img&ei=t6KRYe-

iEIyEieoPtuu_gAY&bih=1041&biw=2133&rlz=1C1CHBF_enUS892US894#imgrc=As2zytXzqAjN_M
[8] Retrieved from https://dashthis.com/analytical-dashboard/
[9] Retrieved from https://ubiq.co/analytics-blog/create-operational-dashboard-business/
[10] Retrieved from *https://www.cascade.app/blog/examples-to-create-strategy-dashboards*
[11] https://www.kmworld.com/Articles/White-Paper/Article/Dashboard-Design-and-Implementation-A-Step-by-Step-Guide-18852.aspx
[12] https://support.sisense.com/kb/en/article/dashboard-planning-and-implementation
[13] *Retrieved from* Pedro Abreu, 2016.Top Things to Learn About Improving Database Performance. Retrieved from https://datacoresystems.ro/index.php/2016/08/

AUTHORS

My Name is Temitope Awodiji, and I am a data Scientist with 10 Years of Experience. I hold a master's degree in Computer Information Science. I am an Efficient Data Analyst and Scientist professional with expert skills in SQL, Power BI, Tableau, EXCEL, and other data analytics tools. My experience includes generating, manipulating, interpreting, and analysing data in a fast-paced delivery and operations.

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

THE FUTURE OF ONLINE LEARNING USING ARTIFICIAL INTELLIGENCE

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China

ABSTRACT

Online learning is the emerging technique in education and learning during the COVID-19 pandemic period. Traditional learning is a complex process as learning patterns, approach, skills and performance varies from person to person. Adaptive online learning focuses on understanding the learner's performance, skills and adapts to it. The use of advanced technology also provides a means to analyse the behavioural learning pattern. As it provides the detailed skill mapping and performance which enables the learner to understand the areas needs to be improved. The information can also be used by assessors to improve the teaching approach. Advanced online learning system using artificial intelligence is an emerging concept in the coming years. In this new concept, the classes are not taken face-to-face in a classroom but through an electronic medium as a substitute. These virtual learning approach are gaining importance every day and very soon they are going to be an integral part of our world. Taking up these virtual learning through an electronic medium is termed as online learning. We proposed two new models which are powered by artificial intelligence (AI) tools. A number of examples of using these new models are presented.

KEYWORDS

Analysis Algorithm, Artificial Intelligence, Hybrid Integrated Model, Online Learning, Progressive Response Learning.

1. INTRODUCTION

The purpose of this adaptive and advanced online learning system is to enable learners to get proper knowledge of course and adjusting system according to the users IQ level [1]. It is an educational method which uses computers as interactive teaching devices and reduces the teacher's workload and enhance the learner understanding. As the online learning system integrated with Artificial Intelligence (AI) is a new approach it has many flaws. Learning is a sophisticated process. And if the computers have to replace a human teacher they need to be more intelligent as Creative Intelligence Learning (CIL) approach does not provide the quantitative teaching learning mechanism [2]. To strengthen the online learning process with the AI mechanism the paper proposed two new adaptive online learning models, 1: Statistical Analysis using AI Learning Model, and 2: Progressive Response Learning Model. This paper will elaborate the models in detail.

2. STATISTICAL ANALYSIS USING ARTIFICIAL INTELLIGENCE LEARNING MODEL

Many online learning assessment systems that use multiple choice approach is based on the correct answers to judge a learner on their understanding of what they have learned [3]. We have

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 187-194, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111915

carry out various experiments on these quantifiable measurements (assessment indicators) on Mathematics and English subject on different learner groups. With these assessment indicators, assessors and learners can easily assess the online learning performances [4].

In our research, we have identify 3 critical assessment indicators which can influence the learners' learning progress and understanding. These 3 assessment indicators have interrelationships with the underlying final mark from the assessment [5]. At the end of the assessment, the artificial intelligence engine will analyse all the statistical information from these 3 indicators and provide a recommendation for the assessor and learner.

(1). **Difficulty Level** (measure by the complexity of the questions [6]). Each question will have the difficulty level embedded. For example, we take a topic in Addition from Mathematics subject. For an Addition topic, we can assign the Difficulty level to these 3 questions depending on the complexity, i.e. 4 + 3 = ? (Low Difficulty), 755 + 958 = ? (Medium Difficulty) and 7,431,398,214 + 32,883,295 = ? (High Difficulty).

Level Terms	Quantitative Measurement
High (hardest)	*3*
Medium	*2*
Low (easiest)	*1*

(2). **Understanding Level** (measure by the time from the question appear to submission). Each question will have the understanding level embedded. For example, assuming there is a question with Level Term - High (from a to b) where "a = 3 seconds" and "b = 5 seconds". In this example, if the learner can submit the answer between 3 to 5 seconds after the question appeared than the answer will be assigned 2 points for the Understanding indicator.

Level Terms	Quantitative Measurement
High (from a to b) fastest	*2*
Medium (from b to c)	*1*
Low (from c onwards) slowest	*0*

(3). **Confident Level** (variation in choosing an answer before submission).

For each question, we will capture the behaviour of the learner when choosing an answer before submission [7]. For example, for most learners if they are confident and prudent on choosing the correct answer, they will submit the answer once decided without making any changes. If the learner didn't make any changes when answer this example question, than this answer will be assigned 2 points for the Confident indicator.

Level Terms	Quantitative Measurement
High (no change on first pick)	*2*
Medium (one change)	*1*
Low (two changes or more)	*0*

In this research, we have carry out multiple experiments to evaluate the use of assessment indicators and the statistical information generated when the learner performing the assessment [8][9]. We have conducted 3 detail experiments and the outcomes generated shows promising result on the learners' overall learning performances. Below are the 3 experiments summary which we have conducted on 100 online voluntary learners. All marks and indicators have been converted to percentage (%) prior for further analysis by the AI engine.

Both the assessment indicators and statistical information stand alone do not have any representations and it is meaningless without others being analysed altogether [10][11]. Furthermore, in order to have an effective and efficient online learning outcome for the learner, the assessor requires to design and develop the curriculum, learning materials and Q&A using a hybrid integrated model. The curriculum needs to be an all rounded learning blueprint, where learner can improve their understanding in a progressive manner and user-friendly approach.

2.1. AI Rules

The AI rules define the way the online learning system assigned learning materials and exercises for the learner to follow. These are the basic rules which we have carry out in our experiments, in which we find it effective in improving the learners understanding. Online learning assessor and learner can modify all the assessment indicators accordingly (depending on various conditions and overall standard requirements) [12].

Rule number	Difficulty level	Correct answers (%)	Understanding level (%)	Confident level (%)	Recommendation (Response)
1	1	< 50	Nil	Nil	Repeat the same difficulty level = 1 exercise
2	1	≥ 50	< 50	< 50	Repeat the same difficulty level = 1 exercise
3	1	≥ 50	< 50	≥ 50	Repeat the same difficulty level = 1 exercise
4	1	≥ 50	≥ 50	< 50	Repeat the same difficulty level = 1 exercise
5	1	≥ 50	≥ 50	≥ 50	Move to next difficulty level = 2 exercise
6	2	< 50	Nil	Nil	Repeat the same difficulty level = 2 exercise
7	2	≥ 50	< 50	< 50	Repeat the same difficulty level = 2 exercise
8	2	≥ 50	< 50	≥ 50	Repeat the same difficulty level = 2 exercise
9	2	≥ 50	≥ 50	< 50	Repeat the same difficulty level = 2 exercise
10	2	≥ 50	≥ 50	≥ 50	Move to next difficulty level = 3 exercise
11	3	< 75	Nil	Nil	Repeat the same difficulty level = 3 exercise
12	3	≥ 75	< 50	< 50	Repeat the same difficulty level = 3 exercise

13	3	≥ 75	< 50	≥ 50	Repeat the same difficulty level = 3 exercise
14	3	≥ 75	≥ 50	< 50	Repeat the same difficulty level = 3 exercise
15	3	≥ 75	≥ 50	≥ 50	Move to next topic exercise

Figure 1. The AI rules applied in the experiments.

3. PROGRESSIVE RESPONSE LEARNING MODEL

In a traditional learning model involves a teacher giving lectures to a group of students in a physical classroom. The teacher will then teach what has been prepared and planned in the curriculum. However, some students have different learning pace and understanding level. This will creates many problems not only to the teacher but to the entire class. Therefore, this Progressive Response Learning Model can min minimise these problems, where students having different learning pace and understanding level.

The principle behind this Progressive Response Learning model is 'to make those fast pace learner go fast while slow pace learner go far'. ln this paper, we will present two analysis where this model can demonstrate the effectiveness when apply to online learning. In these two analysis, I will use multiple choice questions approach and the outcomes from the assessment t will than be evaluated. Afterthat, the assessor and learner can easily review the online learning performance [13).

The analysis consists of 2 assessment groups from 20 voluntary online learners between the age of 9 and 10. We assigned them into 2 groups based on their school assessment results in Algebra topic, Mathematics subject, GROUP (A) - 10 fast pace learners and GROUP (B) - 10 slow pace learners. Each group will be assigned 10 Q&A sections and each Q&A section will have 10 questions. In total, we will apply 100 questions with the Difficulty level distribution of 400/o Low, 30% Medium and 30% High. I n this analysis, we will compare the traditional learning approach and the Progressive Response Leaning approach. The Q&A questions are from Year 5: UK National Curriculum, Algebra topic (Mathematics).

3.1. Traditional Learning Approach

All leaners (fast and slow pace leaners) are require to do all I 0 Q&A sections, each section a day, within 10 days. The outcomes from this assessment are represented in marks (in %) and groups, GROUP (A) - Fast-Pace Learners and GROU P (B) - Slow Pace Learners.

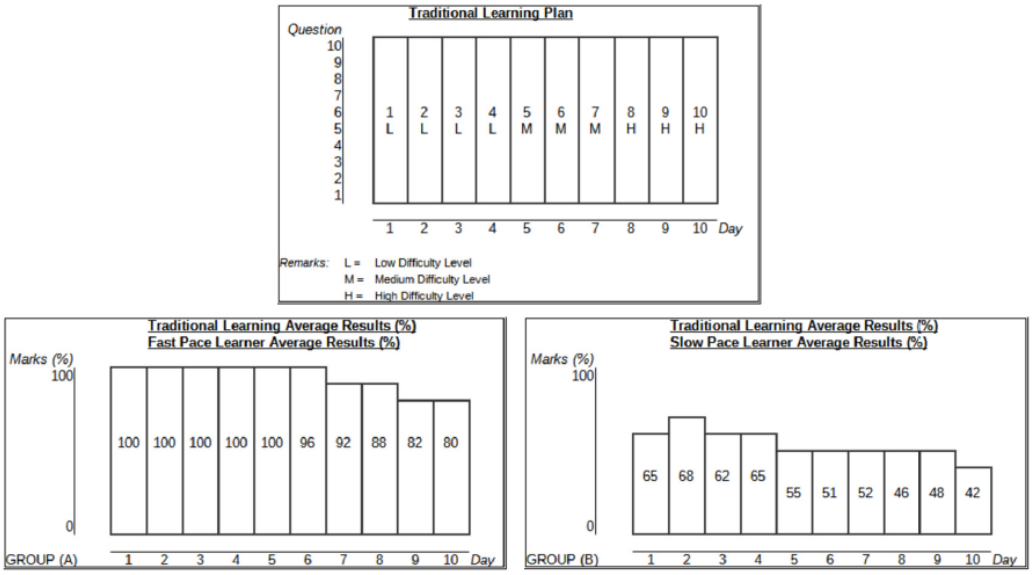


Figure 2. The results from applying traditional learning approach on two different groups of learners, GROUP (A) - Fast Pace Learners group and GROUP (B) -Slow Pace Learners group.

3.2. ProgressiveResponseLearningApproach

All leaners (fast and slow pace learners) are require to do all I 0 Q&A sections, each section a day, within 10 days. In this model, we will repeat some of the previous Q&A in every exercise and using this approach, we can encourage the learner to retrieve the skills the learner learned from the previous exercise and increase the confidence level when facing new quest ions. The outcomes from this assessment are represented in marks (in %) and the Difficulty level are presented on the charts (Figure 3).

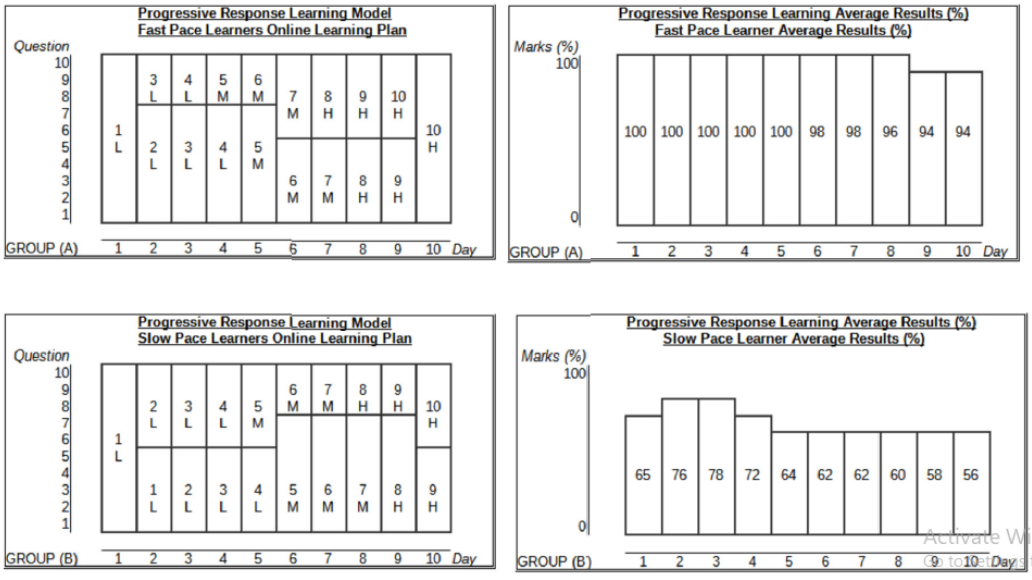


Figure 3. The results from applying Progressive Response Learning approach on twodifferent groups of learners.
GROUP (A)-Fast Pace<Leaners group and GROUP (B) -Slow Pace Learners group.

As we can see from the preliminaryexperimental results, there are improvements in the overall marks and performance after applying the Progressive Response Learning model. Besides that, this model involves psychological factor when facing different learner's behaviour, therefore, further research and investigations are needed in this area.

4. ONLINE LEARNING SYSTEM

The online learning using Al system include several components, whichcan be integrated as one complete artificial intelligence online learning system (14). These are the standard components:·

1. Reasoning - It is the set of processes that empowers us to provide basis for judgement, making decisions, and prediction.
2. Learning - It is the activity of gaining in formation or skill by studying, practising, being educated, or experiencing something. Leaning improves the awareness of the subjects of the study.
3. Problem Solving - It is the procedure in which one perceives and tries to arrive at a desired solution from a current situation by taking some path, which is obstructed by known or unknown hurdles.
4. Perception - It is 1he way of acquiring, interpreting, selecting, and organizing sensory information.
5. Linguistic Intelligence - It is one's ability to use, comprehend, talk, and compose the verbal and written language. It is significant in interpersonal communication.

The potential of online learning system include 4 factors of accessibility, flexibility, interactivity, and collaboration of online learning afforded by the technology. In terms of the challenges to online learning, 6 are identified: defining online learning; proposing a new legacy of epistemology-social constructivism for all; quality assurance and standards; commitment versus innovation; copyright and intellectual property; and personal learning in social constructivism.

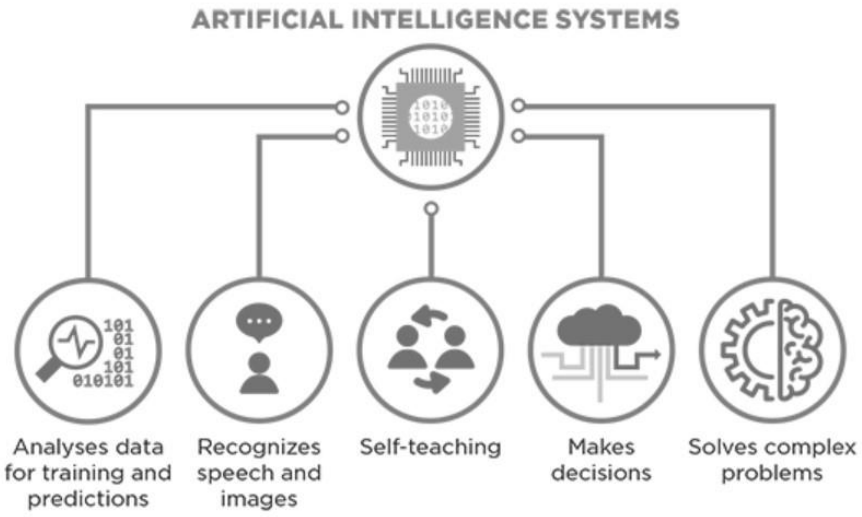


Figure 4. The artificial intelligence online learning system components.

5. CONCLUSIONS

Adaptive online learning is an enhancement that makes learning systems more effective by adapting the presentation of information and overall link structure to the individual learner, based

on learners' knowledge and behaviour. The teachers can use the learner behaviour information for various analyses and make the changes in the teaching process to improve the teaching and learning process. By further expert analysis and experimentation it can become a firm educational method which uses computers as interactive teaching devices to enhance individual skills. It is based on project methodology in which learner's cognitive and psychological will be judged based on learner overall performance. It is based on AI domain and further scope of improvement is huge. The two new proposed models show promising response in AI online learning and further evaluation and research is in progress.

REFERENCES

[1] Jonathan Michael Spector, Du Jing, (2017). Artificial Intelligence and the Future of Education: Big Promises – Bigger Challenges, ACADEMICS, No. 7.

[2] Oscar Sanjuan, B. Cristina Pelayo Garcia-Bustelo, Ruben Gonzalez Crespo, Enrique Daniel France, (2009). Using Recommendation System for E-Learning Environment at degree level, International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 1, No. 2.

[3] S. M. Patil, T. D. Shaikh, (2014). Implementing Adaptability in E-Learning Management System Using Moodle for Campus Environment, International Journal of Emerging Technology and Advanced Engineering, Vol. 4, No. 8.

[4] Shraddha Kande, Pooja Goswami, GurpreetNaul, Mrs. Nirmala Shinde, (2016). Adaptive and Advanced E-learning Using Artificial Intelligence, Journal of Engineering Trends and Applications, Vol. 3, No. 2.

[5] Ofra Walter, VeredShenaar-Golan and Zeevik Greenberg, (2015). Effect of Short-Term Intervention Program on Academic Self-Efficacy in Higher Education, Psychology, Vol. 6, No. 10.

[6] Calum Chace, (2019). Artificial Intelligence and the Two Singularities, Chapman & Hall/CRC.

[7] PieroMella, (2017). Intelligence and Stupidity – The Educational Power of Cipolla's Test and of the "Social Wheel", Creative Education, Vol. 8, No. 15.

[8] Zhongzhi Shi, (2019). Cognitive Machine Learning, International Journal of Intelligence Science, Vol. 9, No. 4.

[9] Crescenzio Gallo and Vito Capozzi, (2019). Feature Selection with Non Linear PCA: A Neural Network Approach, Journal of Applied Mathematics and Physics, Vol. 7, No. 10.

[10] Charles Kivunja, (2015). Creative Engagement of Digital Learners with Gardner's BodilyKinesthetic Intelligence to Enhance Their Critical Thinking, Creative Education, Vol. 6, No. 6.

[11] EvangeliaFoutsitzi, Georgia Papantoniou, EvangeliaKaragiannopoulou, HarilaosZaragas and Despina Moraitou, (2019). The Factor Structure of the Tacit Knowledge Inventory for High School Teachers in a Greek Context.

[12] Nick Bostrom and Eliezer Yudkowsky, (2011). The Ethics of Artificial Intelligence, Cambridge University Press.

[13] Gus Bekdash, (2019). Using Human History, Psychology, and Biology to Make AI Safe for Humans, Chapman & Hall/CRC.

[14] The Student Circles.com, Artificial Intelligence Study Notes https://www.thestudentcircle.com/quickguide.php?url=artificial-intelligence

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, HewlettPackard (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.

MITIGATION TECHNIQUES TO OVERCOME DATA HARM IN MODEL BUILDING FOR ML

Ayse Arslan

Oxford Alumni of Northern California, Santa Clara, USA

ABSTRACT

Given the impact of Machine Learning (ML) on individuals and the society, understanding how harm might be occur throughout the ML life cycle becomes critical more than ever. By offering a framework to determine distinct potential sources of downstream harm in ML pipeline, the paper demonstrates the importance of choices throughout distinct phases of data collection, development, and deployment that extend far beyond just model training. Relevant mitigation techniques are also suggested for being used instead of merely relying on generic notions of what counts as fairness.

KEYWORDS

Fairness in machine learning, societal implications of machine learning, algorithmic bias, AI ethics, allocative harm, representational harm.

1. INTRODUCTION

Algorithms do not "decide" or "guide" or "theorize"—the human-beings developing those algorithms do. Yet, developers and software engineers are often unable to anticipate the consequences that arise when their code and embedded assumptions interact with a complex (and unequal) world, and how that interaction will reinforce (or misguide) our interpretations of human behaviour. Algorithms, in other words, do not only help us parse data; they also generate data that will then be analysed as resulting from human behaviour.

This paper provides a framework for understanding different sources of harm throughout the ML life cycle in order to offer techniques for mitigations based on an understanding of the data generation and development processes rather than relying on generic assumptions of what being fair means.

2. EXISTING WORK

An ML algorithm aims to find patterns in a (usually massive) dataset, and to apply that knowledge to make a prediction about new data points (e.g: photos, job applicant profiles, medical records etc.) (Cusumano et al., 2019; Parker, van Alstyne, & Choudary, 2016). As a result, problems can arise during the data collection, model development, and deployment processes that can lead to different harmful downstream consequences.

This paper refers to the concept of "harm" or "negative consequences" caused by ML systems. ML (Machine Learning) can be defined as the overall process inferring in a statistical way from existing data in order to generalize to new, unseen data.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 195-207, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111916

Deep reinforcement learning—where machines learn by testing the consequences of their actions—combines deep neural networks with reinforcement learning, which together can be trained to achieve goals over many steps. Most machine learning algorithms are good at perceptive tasks such as recognizing a voice or a face. Yet, deep reinforcement learning can learn tactical sequences of actions, things like winning a board game or delivering a package. In the real world, human-beings are able to very quickly parse complex scenes where simultaneously many aspects of common sense related to physics, psychology, language and more are at play.

A high-level overview of a ML-based model might look as follows:

Data Collection

Before any analysis or learning happens, data must first be collected. Compiling a dataset involves identifying a target population (of people or things), as well as defining and measuring features and labels from it. Often, ML practitioners use existing datasets rather than going through the data collection process.

Data Preparation

Depending on the data modality and task, different types of preprocessing may be applied to the dataset before using it.

As Figure 1 displays, the data generation process begins with data collection. This process involves defining a target population and sampling from it, as well as identifying and measuring features and labels. This dataset is split into training and test sets. Data is also collected (perhaps by a different process) into benchmark datasets.

Model Development

Models are then built using the training data (not including the held-out validation data).

As seen in Figure 1, a model is defined, and optimized on the training data. Test and benchmark data is used to evaluate it, and the final model is then integrated into a real-world context. This process is naturally cyclic, and decisions influenced by models affect the state of the world that exists the next time data is collected or decisions are applied. The red color indicate where in this pipeline different sources of downstream harm might arise.

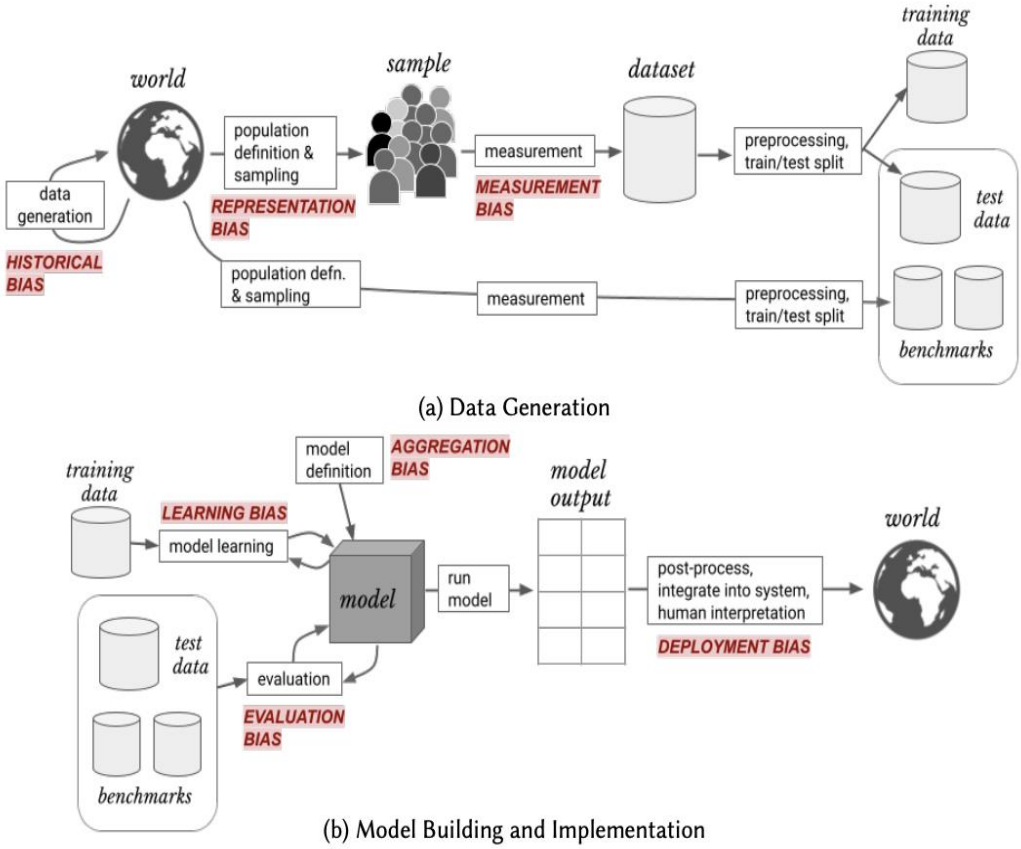


Fig. 1. Overview of ML data generation and model development

When it comes to building models for images, a training dataset can be developed for a particular detection model by manually labelling images. In case of a low precision and recall, for example due to the variety of natural and artificial features, following methods can be utilized to improve performance:

- The use of mix-up as a regularization method, where random training images are blended together by taking a weighted average. Though mix-up is originally proposed for image classification, it can be used for semantic segmentation. Regularization is important in general for segmentation task, as even with 100k training images, the training data might not capture the full variation of terrain, atmospheric and lighting conditions that the model is presented with at test time, and hence, there is a tendency to overfit.

- Another method is the use of unsupervised self-training in which the output of the best detection model from the previous stage is used as a ‘teacher’ to then train a ‘student’ model that makes similar predictions from augmented images. In practice, this could reduce false positives and sharpen the detection output. In order to overcome the issue of “blobby” detections, one can use distance weighting to adapt the loss function for making correct predictions near boundaries. During training, distance weighting places greater emphasis at the edges by adding weight to the loss — particularly where there are instances that nearly touch.

When visually inspecting the detections for low-scoring images, various causes can be noted such as problematic label errors. In order to shed light onto which methods contribute most to the final performance, mean average precision (mAP) can be measured. Distance weighting, mixup and the use of ImageNet pre-training are most common factors for the performance of the supervised learning baseline.

Model Evaluation

After the final model is chosen, the performance of the model on the test data is reported. The test data is not used before this step, to ensure that the model's performance is a true representation of how it performs on unseen data. Aside from the test data, other available datasets — also called benchmark datasets — may be used to demonstrate model robustness or to enable comparison to other existing methods.

Model Post-processing

Once a model has been trained, there are various post-processing steps that may needed. For example, if the output of a model performing binary classification is a probability, but the desired output to display to users is a categorical answer, there remains a choice of what threshold(s) to use to round the probability to a hard classification.

Model Deployment

There are many steps that arise in deploying a model to a real-world setting. For example, the model may need to be changed based on requirements for explainability or apparent consistency of results, or there may need to be built-in mechanisms to integrate real-time feedback. Importantly, there is no guarantee that the population a model sees as input after it is deployed (here, we will refer to this as the use population) looks the same as the population in the development sample.

The algorithms used to parse and analyze those data become commercial black boxes. Barocas et al. [4] provide a useful framework for thinking about how these consequences actually manifest, splitting them into allocative harms (when opportunities or resources are withheld from certain people or groups) and representational harms (when certain people or groups are stigmatized or stereotyped). For example, algorithms that determine whether someone is offered a loan or a job [12, 36] risk inflicting allocative harm. We, human-beings are fallible in making unbiased decisions ourselves and algorithms can actually help us detect human-generated (and socially reinforced) discrimination (Kleinberg et al., 2020; Mullainathan, 2019).

There's a large body of work on testing common sense and reasoning in AI systems. Many of them are focus on natural language understanding, including the famous Turing Test and Winograd schemas. In contrast, the AGENT project focuses on the kinds of reasoning capabilities humans learn before being able to speak. The idea behind the AGENT (Action, Goal, Efficiency, coNstraint, uTility) test by DeepMind Team is to assess how well AI systems can mimic this basic skill, what they can develop psychological reasoning capabilities, and how well the representations they learn generalize to novel situations.

According to the DeepMind Team, the AGENT test takes place in two phases:

- First, the AI is presented with one or two sequences that depict the agent's behavior. These examples should familiarize the AI with the virtual agent's preferences.

- After the familiarization phase, the AI is shown a test sequence and it must determine whether the agent is acting in an expected or surprising manner.

The designers of the tests have included human inductive biases, which means the agents and environment are governed by rules that would be rational to humans (e.g., the cost of jumping or climbing an obstacle grows with its height). This decision helps make the challenges more realistic and easier to evaluate.

The Deepmind researchers tested the AGENT challenge on two baseline AI models. The first one, Bayesian Inverse Planning and Core Knowledge (BIPaCK), is a generative model that integrates physics simulation and planning.

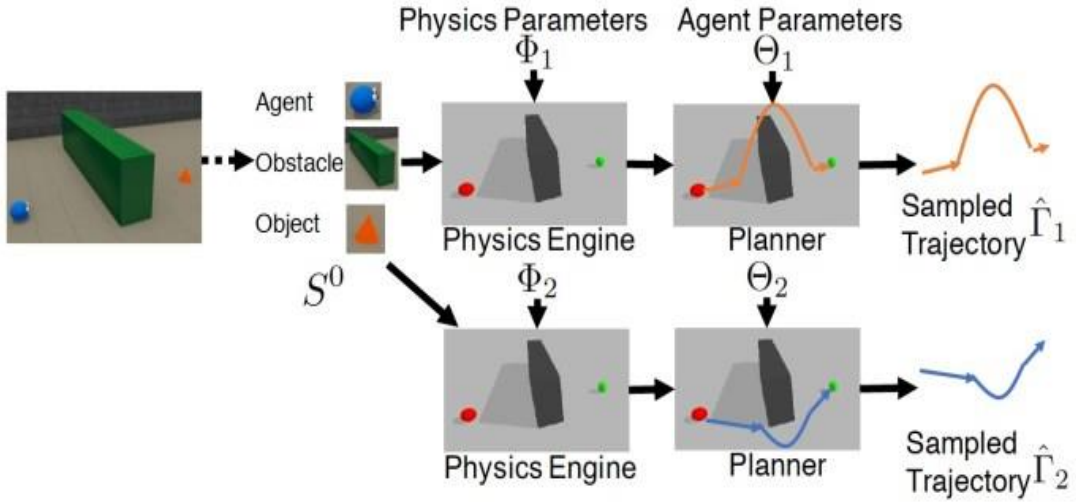


Fig. 2. Overview of BIPaCK Model

As seen Figure 2., the BIPaCK model uses planner and physics engines to predict the trajectory of the agent. The model uses the full ground-truth information provided by the dataset and feeds it into its physics and planning engine to predict the trajectory of the agent. However, in the real world, AI systems don't have access to precisely annotated ground truth information and must perform the complicated task of detecting objects against different backgrounds and lighting conditions.

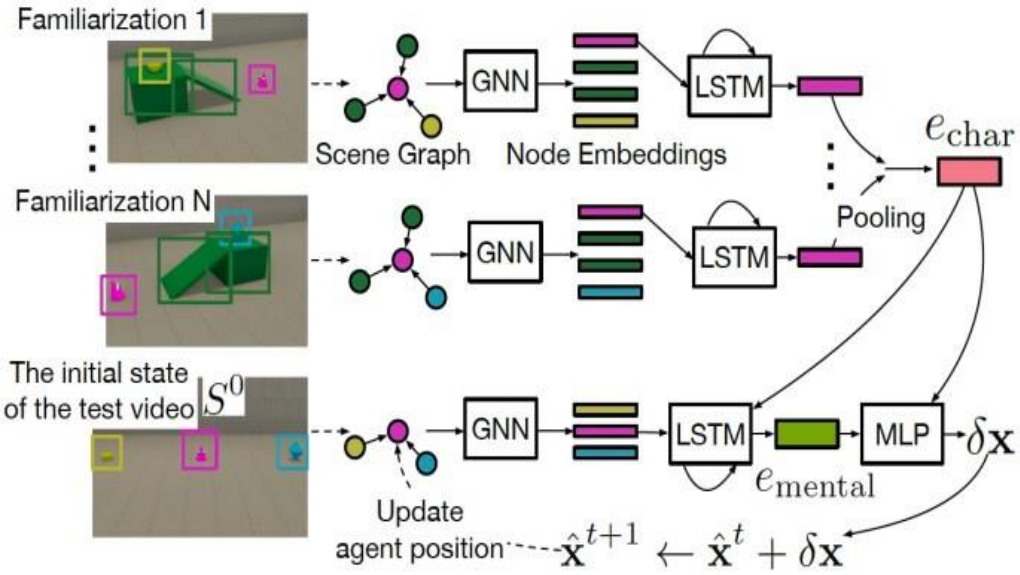


Fig. 3. Overview of ToMnet-G model

The ToMnet-G model uses graph neural networks and LSTMs to embed scene representations and predict agent behavior (Fig. 3). The contrast between the two models highlights the challenges of the simplest tasks that humans learn without any instructions.

In order for an ML model to work well, the following simple steps can be implemented:

1. Train a classifier on labeled data.
2. The bigger classifier model then infers pseudo-labels on a much larger unlabeled dataset.
3. Then, it trains a larger classifier on the combined labeled and pseudo-labeled data, while also adding noise.
4. (Optional) Going back to step 2, the smaller model may be used a new classifier.

One can view this as a form of self-training, because the model generates pseudo-labels with which it retrains itself to improve performance. One underpinning hypothesis is that the noise added during training not only helps with the learning, but also makes the model more robust. This approach is similar to knowledge distillation, which is a process of transferring knowledge from a large model to a smaller model. The goal of distillation is to improve speed in order to build a model that is fast to run in production without sacrificing much in quality compared to the bigger model.

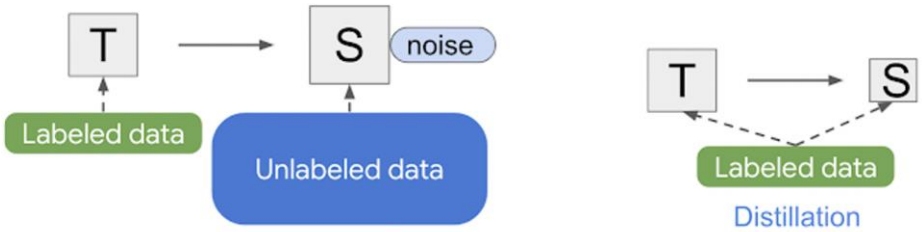


Fig. 4. Simple illustrations of the model and knowledge distillation.

Knowledge distillation does not add noise during training (e.g., data augmentation or model regularization) and typically involves a smaller inference model. In contrast, one can think of it as the process of "knowledge expansion". One strategy for training production models is to apply training twice (Fig. 4):

- first to get a larger inference model T' and then
- to derive a *smaller* model S.

In some cases, the training may need data augmentation, yet, in certain applications, e.g., natural language processing, such types of input noise are not readily available. For those applications, the training model can be simplified to have no noise. In that case, the above two-stage process becomes a simpler method:

- First, the bigger model infers pseudo-labels on the unlabeled dataset from which is a new model (T') that is of *equal-or-larger* size than the original model being trained.
- The self-training phase is then followed by knowledge distillation to produce a smaller model for production.

3. SOURCES OF HARM IN ML

This section explores each potential source of harm in-depth. Each subsection will detail where and how in the ML pipeline problems might arise, as well as a characteristic example. These categories are not mutually exclusive; however, identifying and characterizing each one as distinct makes them less confusing and easier to tackle.

3.1. Historical Bias

Historical bias arises even if data is perfectly measured and sampled, if the world as it is or was leads to a model that produces harmful outcomes. Such a system, even if it reflects the world accurately, can still inflict harm on a population. Considerations of historical bias often involve evaluating the representational harm (such as reinforcing a stereotype) to a particular group.

3.2. Representation Bias

Representation bias occurs when the development sample under-represents some part of the population, and subsequently fails to generalize well for a subset of the use population. Representation bias can arise in several ways:

(1) When defining the target population, if it does not reflect the use population. Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis.
(2) When defining the target population, if contains under-represented groups. Say the target population for a particular medical dataset is defined to be adults aged 18-40. There are minority groups within this population: for example, people who are pregnant may make up only 5% of the target population.
(3) When sampling from the target population, if the sampling method is limited or uneven. For example, the target population for modeling an infectious disease might be all adults, but medical data may be available only for the sample of people who were considered serious enough to bring in for further screening. As a result, the development sample will represent a skewed subset of the target population. In statistics, this is typically referred to as sampling bias.

3.3. Measurement Bias

Measurement bias occurs when choosing, collecting, or computing features and labels to use in a prediction problem. For example, "creditworthiness" is an abstract construct that is often operationalized with a measureable proxy like a credit score. Proxies become problematic when they are poor reflections or the target construct and/or are generated differently across groups, which can happen when:

(1) The proxy is an oversimplification of a more complex construct. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Algorithm designers may resort to a single available label such as "GPA" [28], which ignores different indicators of success present in different parts of the population.
(2) The method of measurement varies across groups. For example, consider factory workers at several different locations who are monitored to count the number of errors that occur (i.e., observed number of errors is being used as a proxy for work quality). This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes [5, 17].
(3) The accuracy of measurement varies across groups. For example, in medical applications, "diagnosed with condition X" is often used as a proxy for "has condition X." However, structural discrimination can lead to systematically higher rates of misdiagnosis or underdiagnosis in certain groups [23, 32, 35].

3.4. Aggregation Bias

A particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different across them. Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (e.g., if there is also representation bias).

3.5. Learning Bias

Learning bias arises when modeling choices amplify performance disparities across different examples in the data [24]. For example, an important modeling choice is the objective function that an ML algorithm learns to optimize during training. Typically, these functions encode some measure of accuracy on the task (e.g., cross-entropy loss for classification problems or mean squared error for regression problems).

3.6. Evaluation Bias

Evaluation bias occurs when the benchmark data used for a particular task does not represent the use population. Evaluation bias ultimately arises because of a desire to quantitatively compare models against each other. Such generalizations are often not statistically valid [38], and can lead to overfitting to a particular benchmark.

3.7. Deployment Bias

Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used. This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated socio-technical system moderated by institutional structures and human decision-makers (Selbst et al. [39] refers to this as the "framing trap").

4. A FRAMEWORK FOR DATA GENERATION AND ML PIPELINE

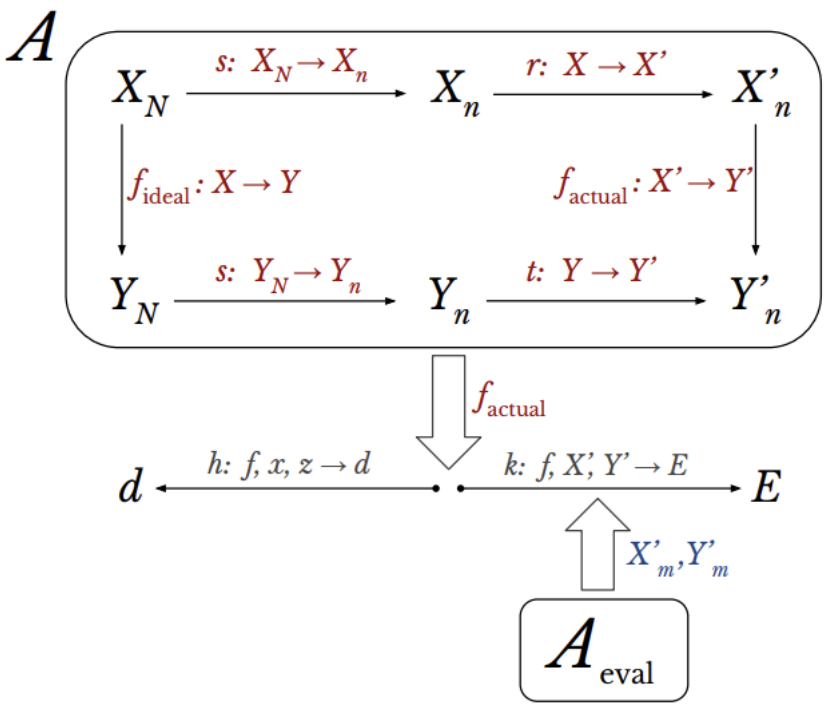


Fig. 5. A data generation and ML pipeline viewed as a series of mapping functions.

There is a growing body of work on "fairness-aware algorithms" that modify some part of the modelling pipeline to satisfy particular notions of "fairness." Consider the data transformations for a dataset as depicted in Figure 5. The upper part of the diagram of Figure 2 deals with data collection and model building, while the bottom half describes the evaluation and deployment process.

The data transformation sequence can be abstracted into a general process A. Let X and Y be the underlying feature and label constructs we wish to capture where $s : X_N \rightarrow X_n$ is the sampling function. X' and Y' are the measured feature and label proxies that are chosen to build a model, where r and t are the projections from constructs to proxies, i.e., $X \rightarrow X'$ and $Y \rightarrow Y'$.

The function $f\text{ideal} : X \rightarrow Y$ is the target function—learned using the ideal constructs from the target population—but $f\text{actual} : X' \rightarrow Y'$ is the actual function that is learned using proxies measured from the development sample. Then, the function k computes some evaluation metric(s) E for $f\text{actual}$ on data X'_m, Y'_m (possibly generated by a different process, e.g., A_{eval} in Figure 2).

Finally, given the learned function $f\text{actual}$, a new input example x, and any external, environmental information z, a function h governs the real-world decision d that will be made (e.g., a human decision-maker taking a model's prediction and making a final decision).

5. MITIGATION TECHNIQUES

The aim of this section is to understand and motivate mitigation techniques in terms of their ability to target different sources of harm to get a better understanding when and why different approaches might help, and what hidden assumptions they make.

As an example, measurement bias is related to how features and labels are generated (i.e., how r and t are instantiated). Historical bias is defined by inherent problems with the distribution of X and/or Y across the entire population. Therefore, solutions that try to adjust s by collecting more data (that then undergoes the same transformation to X') will likely be ineffective for either of these issues. However, it may be possible to combat historical bias by designing s to systematically over- or under-sample X and Y, leading to a development sample with a different distribution that does not reflect the same undesirable historical biases. In the case of measurement bias, changing r and t through more thoughtful, context-aware measurement or annotation processes (e.g., as in Patton et al. [34]) may work.

In contrast, representation bias stems either from the target population definition (X_N , Y_N) or the sampling function (s). In this case, methods that adjust r or t (e.g., choosing different features or labels) or g (e.g., changing the objective function) may be misguided. Importantly, solutions that do address representation bias by adjusting s implicitly assume that r and t are acceptable and that therefore, improving s will mitigate the harm.

Learning bias is an issue with the way f is optimized, and mitigations should target the defined objective(s) and learning process [24]. In addition, some sources of harm are connected: e.g., learning bias can exacerbate performance disparities on under-represented groups, so changing s to more equally represent different groups/examples could also help prevent it.

Deployment bias arises when h introduces unexpected behaviour affecting the final decision d. Dealing with deployment bias is challenging since the function h is usually determined by complex real-world institutions or human decision-makers. Mitigating deployment bias might involve instituting a system of checks and balances in which users balance their faith in model predictions with other information and judgements [26]. This might be facilitated by choosing an f that is human-interpretable, or by developing interfaces that help users understand model uncertainty and how predictions should be used. Evaluation and aggregation bias are discussed in more detail below.

To supplement user reporting, platforms have algorithms that flag content for human review. Several platforms currently use image recognition tools and natural language processing classifiers to help moderators filter and prioritize possible objectionable content for evaluation.

Such prompts have at least three virtues. First, they may help users pause and engage in what Daniel Kahneman calls “system 2” thinking—higher-level cognitive reflection. In fact, research has shown that pop-up warnings requiring user interaction to dismiss them can positively change user behavior. Second, if such self-moderation occurs, it would be in advance of posting, before potentially harmful material can spread. Finally, these prompts preserve users’ freedom of expression, as they allow users to ignore the warnings and post the questionable material anyway.

Finally, there is a risk of exploitation by bad actors. Those who intentionally and willfully post misleading or dangerous material will not be deterred by an algorithmic warning. Instead, they could use the warnings to help them craft harmful posts that fall just below the threshold of algorithmic detection.

User prompts are designed to reduce the spread of harmful content while respecting freedom of expression and are immediate and reasonably effective. The precedent for user prompts already exists, and the technology needed to expand them into new contexts is available. All that remains is for platforms to take action.

6. RECOMMENDATIONS

Bringing convolutional neural networks (CNNs) to any industry through means of AI algorithms—whether it be medical imaging, robotics, or some other application entirely—has the potential to enable new functionalities and reduce the compute requirements for existing processes as a single CNN can replace more computationally expensive image processing, denoising, and object detection algorithms. However, there might be some challenges and difficulties while moving an idea from conception to productization. Here is an overview of some challenges and potential solutions regarding the development and deployment of AI model.

Leverage existing models

As existing models already exist for almost every application, rather than reinventing the wheel, it's often much easier to start with a network based on one of these architectures. Moreover, starting with a known model will reduce the amount of time, data, and effort to train a model, since it's possible to retrain existing models in a process called 'transfer learning.'

Simple models are effective

For most applications, there is no need for a latest and greatest in CNN architectures. For example, if an application only requires detecting the difference between a few different objects with high certainty, even simple detectors can do the task. Users can benefit greatly once they realize that their applications can be solved for a fraction of the computational complexity with much simpler models than what's on the forefront of research. The goal is to not make the migration to CNNs any harder than it has to be.

Integrate quantization early

Quantizing a model down from multi-byte precisions to a single-byte can multiply inference speed with little to no degradation in accuracy. For example, frameworks such as PyTorch expose their own methods for quantizing models, but they're not always compatible with each other. Regardless of the approach taken, the aim should be to quantize from the outset of developing the model in a consistent way.

7. CONCLUSION

This paper provides a framework for understanding the sources of downstream harm caused by ML systems to facilitate productive communication around potential issues. By framing sources of downstream harm through the data generation, model building, evaluation, and deployment processes, we encourage application-appropriate solutions rather than relying on broad notions of what is fair. Fairness is not one-size-fits-all; knowledge of an application and engagement with its stakeholders should inform the identification of these sources.

Finally, the paper illustrates that there are important choices being made throughout the broader data generation and ML pipeline that extend far beyond just model training. In practice, ML is an iterative process with a long and complicated feedback loop. This paper highlighted problems

that manifest through this loop, from historical context to the process of benchmarking models to their final integration into real-world processes.

REFERENCES

[1] Agre, P. E. (1994). Surveillance and capture: Two models of privacy. The Information Society, 10(2), 101–127.
[2] Allen, J. (2016). Topologies of power. Beyond territory and networks. Routledge.
[3] Bratton, B. (2015). The Stack: On software and sovereignty. MIT Press.
[4] Bucher, T. (2018). If...then: Algorithmic power and politics. Oxford University Press.
[5] Castañeda, L., & Selwyn, N. (2018). More than tools? Making sense of the ongoing digitizations of higher education. International Journal of Educational Technology in Higher Education, 15(1).
[6] Decuypere, M. (2019a). Open Education platforms: Theoretical ideas, digital operations and the figure of the open learner. European Educational Research Journal, 18(4), 439–460.
[7] Decuypere, M. (2019b). Researching educational apps: ecologies, technologies, subjectivities and learning regimes. Learning, Media and Technology, 44(4), 414–429.
[8] Decuypere, M. (2019c). STS in/as education: where do we stand and what is there (still) to gain? Some outlines for a future research agenda. Discourse: Studies in the Cultural Politics of Education, 40(1), 136–145
[9] Dieter, M., Gerlitz, C., Helmond, A., Tkacz, N., Vlist, F., Der, V., & Weltevrede, E. (2018). Store, interface, package, connection : Methods and propositions for multi-situated app studies. CRC Media of Cooperation Working Paper Series No 4.
[10] Drucker, J. (2020). Visualization and Interpretation: Humanistic Approaches to Display. MIT Press. Journal of New Approaches in Educational Research, 10(1)
[11] Mathias, Decuypere The Topologies of Data Practices: A Methodological Introduction Fedorova, K. (2020). Tactics of Interfacing. Encoding Affect in Art and Technology. MIT Press. Goriunova, O. (2019). The Digital Subject: People as Data as Persons. Theory, Culture & Society, 36(6), 125–145.
[12] & Ruppert, E. (2020). Population Geometries of Europe: The Topologies of Data Cubes and Grids. Science, Technology, & Human Values, 45(2), 235–261.
[13] Gulson, K. N., Lewis, S., Lingard, B., Lubienski, C., Takayama, K., & Webb, P. T. (2017). Policy mobilities and methodology: a proposition for inventive methods in education policy studies. Critical Studies in Education, 58(2), 224–241.
[14] Gulson, K. N., & Sellar, S. (2019). Emerging data infrastructures and the new topologies of education policy. Environment and Planning D: Society and Space, 37, 350–366.
[15] Hartong, S. (2020). The power of relation-making: insights into the production and operation of digital school performance platforms in the US. Critical Studies in Education, 00(00), 1–16.
[16] Hartong, S., & Förschler, A. (2019). Opening the black box of data-based school monitoring: Data infrastructures, flows and practices in state education agencies. Big Data & Society, 6(1),
[17] Lash, S. (2012). Deforming the Figure: Topology and the Social Imaginary. Theory, Culture & Society, 29(4-5), 261–287.
[18] Latour, B. (1986). Visualization and cognition: Thinking with eyes and hands. Knowledge & Society, 6, 1–40. Retrieved from http://hci.ucsd.edu/10/readings/Latour(1986).pdf
[19] Law, J. (2004). After Method: Mess in Social Science Research. Psychology Press.
[20] Lewis, S. (2020). Providing a platform for "what works": Platform-based governance and the reshaping of teacher learning through the OECD's PISA4U. Comparative Education, 56(4).
[21] Lewis, S., & Hardy, I. (2017). Tracking the Topological: The Effects of Standardised Data Upon Teachers' Practice. British Journal of Educational Studies, 65(2), 219–238.
[22] Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. New Media and Society, 20(3), 881–900.
[23] Lindh, M., & Nolin, J. (2016). Information We Collect: Surveillance and Privacy in the Implementation of Google Apps for Education. European Educational Research Journal, 15(6),
Lury, C., & Day, S. (2019). Algorithmic Personalization as a Mode of Individuation. Theory, Culture & Society, 36(2), 17–37.

[24] Mathias, Decuypere The Topologies of Data Practices: A Methodological Introduction Lury, C., Fensham, R., Heller-Nicholas, A., & Lammes, S. (2018). Routledge Handbook of Interdisciplinary Research Methods. Routledge.
[25] Lury, C., Parisi, L., & Terranova, T. (2012). Introduction: The Becoming Topological of Culture. Theory, Culture & Society, 29(4-5), 3–35.
[26] Lury, C., Tironi, M., & Bernasconi, R. (2020). The Social Life of Methods as Epistemic Objects: Interview with Celia Lury. Diseña, 16, 32–55.
[27] Lury, C., & Wakeford, N. (2012). Introduction: A perpetual inventory. Inventive Methods (pp. 15–38). Routledge.
[28] Martin, L., & Secor, A. J. (2014). Towards a post-mathematical topology. Progress in Human Geography, 38(3), 420–438.
[29] Piattoeva, N., & Saari, A. (2020). Rubbing against data infrastructure(s): methodological explorations on working with(in) the impossibility of exteriority. Journal of Education Policy, 00(00), 1–21.
[30] Plantin, J. C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. New Media and Society, 20(1), 293–310.
[31] Prince, R. (2017). Local or global policy? Thinking about policy mobility with assemblage and topology. Area, 49(3), 335–341.
[32] Ratner, H. (2019). Topologies of Organization: Space in Continuous Deformation. Organization Studies, 1–18.
[33] Ratner, H., & Gad, C. (2019). Data warehousing organization: Infrastructural experimentation with educational governance. Organization, 26(4), 537–552.
[34] Ratner, H., & Ruppert, E. (2019). Producing and projecting data: Aesthetic practices of government data portals. Big Data & Society, 6(2), 1–16.
[35] Ruppert, E., Law, J., & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. Theory, Culture & Society, 30(4), 22–46.
[36] Suchman, L. (2012). Configuration. In C. Lury & N. Wakeford (Eds.), Inventive Methods: The Happening of the Social (pp. 48–60). Taylor and Francis.
[37] Thompson, G., & Cook, I. (2015). Becoming-topologies of education: deformations, networks and the database effect. Discourse: Studies in the Cultural Politics of Education, 36(5), 732–748.
[38] Thompson, G., & Sellar, S. (2018). Datafication, testing events and the outside of thought. Learning, Media and Technology, 43(2), 139–151.
[39] van de Oudeweetering, K., & Decuypere, M. (2019). Understanding openness through (in)visible platform boundaries: a topological study on MOOCs as multiplexes of spaces and times. International Journal of Educational Technology in Higher Education, 16(1).
[40] van de Oudeweetering, K., & Decuypere, M. (2020). In between hyperboles: forms and formations in Open Education. Learning, Media and Technology, Advance online publication, 1–18.
[41] Williamson, B. (2017). Learning in the "platform society": Disassembling an educational data assemblage. Research in Education, 98(1), 59–82.

AUTHORS

Ayse received her MSc in Internet Studies in University of Oxford in 2006. She participated in various research projects for UN, Nato and the EU regarding HCI (human-computer interaction). She completed her doctorate degree in user experience design in Oxford while working as an adjunct faculty member at Bogazici University in her home town Istanbul. Ayse has also a degree in Tech Policy from Cambridge University. Currently, Ayse lives in Silicon Valley where she works as a visiting scholar for Google on human-computer interaction design.

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

Unsupervised Named Entity Recognition for Hi-Tech domain

Abinaya Govindan, Gyan Ranjan, and Amit Verma

Neuron7.ai, USA

Abstract. This paper presents named entity recognition as a multi-answer QA task combined with contextual natural-language-inference based noise reduction. This method allows us to use pre-trained models that have been trained for certain downstream tasks to generate unsupervised data, reducing the need for manual annotation to create named entity tags with tokens. For each entity, we provide a unique context, such as entity types, definitions, questions and a few empirical rules along with the target text to train a named entity model for the domain of our interest. This formulation (a) allows the system to jointly learn NER-specific features from the datasets provided, and (b) can extract multiple NER-specific features, thereby boosting the performance of existing NER models (c) provides business-contextualized definitions to reduce ambiguity among similar entities. We conducted numerous tests to determine the quality of the created data, and we find that this method of data generation allows us to obtain clean, noise-free data with minimal effort and time. This approach has been demonstrated to be successful in extracting named entities, which are then used in subsequent components.

Keywords: natural language processing,named entity recognition unstructured data generation, question answering, information retrieval

1 Introduction

The increasing availability of open source Natural Language Processing (NLP) resources and toolkits, combined with the massive amount of data generated every day, necessitates the development of tools that can analyse this data and extract useful information. Unfortunately, just because there is more data being generated every day does not indicate that it can be used to train modern deep learning systems.

In NLP, named entity recognition (NER) is a crucial task which aims to recognise and classify named items such as persons, locations, and events. These extracted named entities are used in a variety of NLP operations to help make better sense of unstructured data.

Some of the early applications of NER included human name identification in a given system such as [1], question answering models [2] that use entity recognition to improve search results and document summarization systems such as [3], where NER identifies significant parts of text that contribute to summaries.

Neural network-based models have recently improved the performance of NER tasks, due to the advances in deep learning. For various human languages, including English, French, and Chinese, named entity models have been developed.

David C. Wyld et al. (Eds): NLPTA, EDU, DSA, IoTE, VLSI, DPPR, ACITY, AIAA, CNDC - 2021
pp. 209-220, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111917

Since NER is becoming more important across many fields and businesses, domain-specific NER technologies have become the new focus. Several NER models for the medical domain system such as [4], have been created to identify a variety of medical categories, including Genes, chemicals, diseases, and so on. This is due to (a) the availability of open source data for all of these areas, such as [5], (b) These datasets do not have any confidentiality associated and thus suitable for training.

However in some domains, using these architectures may be insufficient because the performance of these models is dependent on the quantity and quality of labeled data, and annotated data generation might be particularly difficult because these models require a large amount of high-quality data. This drives researchers to hope to develop a mechanism for extracting semantic and lexical knowledge from enormous amounts of unstructured, unlabeled data, which can then be applied to the NER task thereby improving the performance.

[6] are creating a Service Intelligence platform that, given a faulty hi-tech hardware, recommends actions and provides actionable insights to the repair technician. While there are commercial applications to create, edit, and search technician notes, historical technician repair notes have not been leveraged to derive insights. A key characteristic of insight recommendation engine is to understand the context of these notes, which is provided by the named entities, and given the unstructured nature of these notes, is not trivial to extract. These extracted named entities a) directly assist technicians by giving focused and most informative segments of the notes, allowing them to spend less time reading and perceiving the notes. b) provide an overview of the problem along with recommendations for parts or locations that the technician should investigate further.

So after careful evaluation based on the above criteria, classes such as *Model name, Parts replaced, Error codes, Frequency, Amplitude, Functional Test performed* are the entity tags that we chose to extract and train.

The goal of this research is to present a system for generating labeled data that can be used directly to train a domain-specific NER model. We perform our experiments on technical case descriptions and technician notes that have been raised on the service intelligence platform. These notes along with the extracted entities provide a technical insight onto what could be the reason behind the exhibited symptom and subsequently the proposed resolution.

From the generated data, we only retain the best quality data, rejecting the ambiguous data points and reducing the noise by employing an ensemble of numerous definitions and business contextualised rules. We also go over the results of fine-tuned models and architectures trained on the generated data.

In this paper, we study various approaches of data generation to train custom named entity modlels. We show our proposed system and novel modules implemented to achieve unsupervised data generation. We also compare the performance

of various deep learning models trained on the generated data along with associated results which depict the effectiveness of our system.

1.1 Related work

Due to the domain specific terminologies and ambiguity of these business terms, Named Entity Recognition (NER) has been deemed a relatively challenging problem in the domain of Hi-Tech. Experiments conducted by [7] show that external knowledge can be helpful in classifying the same terms as different named entities depending on the context. Earlier works like [8] have used rule based and dictionary based approaches to solve the NER task for particular domains and languages. However, these approaches have weak generalisation properties when applied to unseen data.

Other learning based systems developed by [9] abd [10] have wide applications across a variety of domains but have some limitations, such as the lack of specific domain knowledge integration and the inability to handle novel entity types with limited data availability. They are also occasionally under-optimized for accuracy due to the usage of less powerful models, resulting in poor performance in downstream activities. For the task of NER, researchers such as [11] , [12] have developed machine learning models such as Hidden Markov Models (HMM) and conditional random fields (CRF). These machine learning methods, however, demand comprehensive and time-consuming feature identification and extraction, which can be costly in terms of manual labour.

Modern deep learning architectures for NER, such as [13], overcome the issues faced by these models. These models generally function best on massive amounts of labeled data, and are thus frequently created for open-source or academic sources. [14] attempts to solve the paucity of data in a few areas by transferring an ANN model trained on a large labeled dataset to a smaller unlabeled dataset. Other synthetic data generation-based methods include (a) back-translation, noise reduction, and parallel sentence extraction as suggested by [15] (b) data labelling modules that use open source websites such as Wikipedia and Google to label data automatically as proposed by [16].

2 Our work

2.1 Task formulation

NER is defined in traditional systems as a token level multi-class classification task. For the purposes of this study, we will concentrate on data generation and transformation into a format that comprises of text tokens and corresponding named entity tags. The named entity tag is created as

$$B - e_k,\ I - e_k\, and\, O$$

where e_k is the entity type. As a result, given any token and context $C_1, C_2, C_3..C_n$, the unsupervised data generation module generates tags $L_1, L_2, L_3...L_n$ which are used to fine-tune a named entity classifier. The tags must be grouped into B and I for us to identify the beginning and end of each entity. We employ four forms of knowledge context based on our experiments: entity types, questions that represent each entity type, definitions for each entity type, and business rules created for each entity type. Due to the lack of innate business domain features and standards, text sequence by itself may not enable us obtain the optimum labels for the provided data, therefore these knowledge contexts are essential to enhance the quality of the data generated.

2.2 Dataset

The data we used for this study was our service intelligence platform's unstructured agent notes and issue descriptions created by technicians. We must be able to give high-quality named entities with a small margin of error because the extracted entities directly assist the technicians in understanding the problem and prescribing solutions. However, due to many limitations in these unstructured notes, such as domain specialised language, a lack of correct linguistic structure, and shorter chunks of text with condensed information, this is not a simple process.

Due to these restrictions, annotated named entities are required to capture all of these information in the tagged data so that the NER model can learn them. To produce named entities from technician notes for diverse product lines, we use various unsupervised approaches such as question answering, natural language inference, and conditional text generation. Not all entity tags have a similar distribution of tagged data, resulting in an imbalance in the data. We accommodate for this imbalance by giving tailored context along with the text sequence to assist the model to learn these features better.

2.3 Candidate generation

The overall approach for data generation is depicted in 1

The first stage of data generation entails extracting candidate named entities using an ensemble of multiple variants of QA models fine tuned on SQuAD [17].We employ an ensemble because we want to minimise the error that a single model can produce, and extract candidates with high confidence using majority voting approaches. This would allow us to reduce the noise produced by these separate models while also maximising their aggregate performance.For each of the mentioned entities, we create a few templates such as:

- *What were the parts replaced* for **Replaced parts** ,
- *What was the error code mentioned in the note* for **Error code**

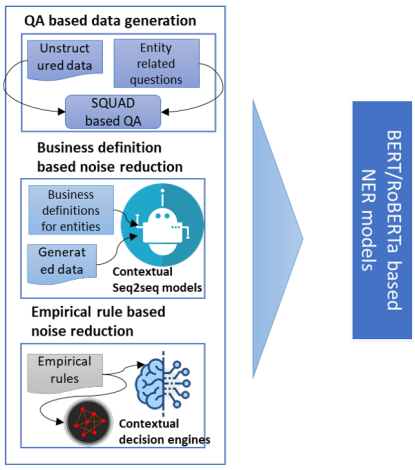


Fig. 1. Architecture for data generation and NER training

– *What was the value of frequency mentioned* for **Frequency**
– *What was the value of amplitude mentioned* for **Amplitude**

During our investigation of the coverage of annotated labels, we determined that a one-to-one question-to-entity type mapping effectively covered these entity tags. For the entity type **Functional test**, we ask a few different variations of questions, such as *What test was performed , What was the name of the test that failed etc.* . These question variants enabled us capture a large number of annotated sample for this entity type. In this dataset, these were the best-performing questions during inference.

As demonstrated in [18], we used several models fine tuned on SQuAD2.0 dataset. Using these models in an ensembled fashion helps in more accurate named entities when compared to using a single model, which in turn improves the final NER model's performance.

2.4 Context and definition based candidate noise reduction

Once we've arrived at a list of candidate entities, we finalised a set of carefully crafted definitions for each of the entity tag, based on business expertise that was provided by the technicians and experts. The definitions were

– Frequency is the rate at which current changes direction per second. Frequency is measured in Hertz (Hz)
– Amplitude is the maximum displacement or distance moved by a point on a vibrating body or wave measured from its equilibrium position. I. Amplitude is measured in decibels (dBs)

– A functional test refers to an operational test performed on a machine that indicates which part of a machine fails.

We incorporate these definitions for additional noise reduction by providing this as a context to a text generation model which was fine tuned to answer Boolean questions as in [19]. To validate the predicted entity for each entity type, we provide the curated entity type definition as additional context in the format "**Context(definition) - Sequence**" as input to the boolq based text generation model. We also provide question to this model in the format, "**Does e_k follow the context mentioned?**" where e_k is the predicted entity. The fine-tuned model responds with either a "yes" or "no" answer, allowing us to retain the candidates which confine to the business contexts.

Since the fine tuned QA ensemble isn't powerful enough to provide only the appropriate entities due to absence of domain knowledge, and is frequently linked with false positives, this textual generation-based noise reduction helps us increase the coverage of predicted entities with minimum false positive rate.

2.5 Business rule based Candidate enrichment

In addition to the entities generated by the question answering models, we created a set of business rules to extract a few other entity types that could not be processed by question answering approaches due to lack of linguistic structure in the notes in which these entities appear. These entities include named entities that we have some prior knowledge about and would like the model to learn for generalisation purposes, such as Model Number. We generate annotated data for these labels using a collection of "seed entities" as the knowledge context. When we utilise this data to fine-tune the model, it learns the latent patterns in which these entities appear and predicts newer entities of this type that may appear in the future.

2.6 Aggregation of generated data

Once the data has been generated using these approaches, we pass it to a collator, which aggregates all of the various predictions in such a way that

– Each token has been tagged to only one single entity, since in our use case, it has been proven that there would be no scenario where one token would belong to more than one entity type.
– Each text sequence has a set of non-overlapping entities. The data generation pipeline tags multiple sub sequences of text to a particular entity type, there are cases where they might end up overlapping with each other. For example, in the sequence *The machine failed self-test during boot up* may have sequences such as *failed self-test , failed self-test during boot up* tagged to the entity type **Functional tests**. The collator looks at various factors such as model prediction

probability, length of sequence tagged and linguistic properties such as noun chunks to choose the most relevant sub sequence among these sequences.

2.7 Explanation and breakup of tagged data numbers

At the end of this module, for the unstructured agent note inputs, we have credible tagged data which can now be fed into our named entity fine tuner. The distribution of the tagged data is as depicted in 1 :

Table 1. Data distribution

Entity type	Number of samples	Percentage of contribution
O	111662	79.69%
I-Test	17565	12.53%
B-Test	5539	3.95%
I-Replaced Parts	1895	1.35%
B-Replaced Parts	1057	0.75%
B-model Number	624	0.45%
I-Amplitude	503	0.36%
B-Frequency	474	0.34%
B-Amplitude	397	0.28%
I-Frequency	334	0.24%
B-error code	35	0.02%
I-error code	31	0.02%
I-model Number	13	0.01%

3 Model Description

3.1 Model Architecture

The basic architecture we use for model training is depicted in 2

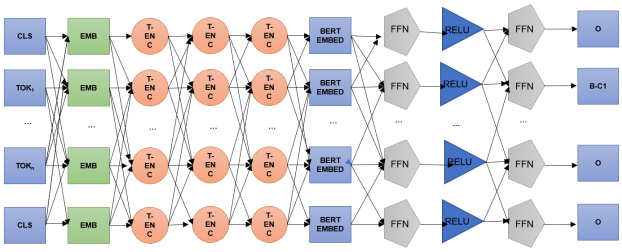


Fig. 2. BERT/ELECTRA/RoBERTa based NER model architecture

Based on the experiments we performed, we chose to fine-tune our NER task on pre-trained architecture such as BERT, RoBERTa and ELECTRA and compare

the performance that was best suited for our use case. From the generated data, we provide tokens in the sequence and named entity types with the B and I tags (to indicate if the token is a start or rest of the named entity) as input to the model. So we define a set of tokens $C = \{c_j\}$ i.e. $[CLS], c_1, c_2, ..., c_n [SEP]$ as input to the pretrained model where n is the number of tokens present in the text. We provide named entity tags such as $O, B - e_k, I - e_k, ..., O, O$ as output for our model. For the input text, the model predicts a tag for each token at the end of the output layer.

3.2 Contextualization

We chose transformer based architectures such as BERT, RoBERTa and ELECTRA to find the ideal model for our use case. We modified the architectures by adding a contextualisation layer in terms of two fully connected layers. We chose fully connected layers since a we intend to learn features from all the combinations of the embedding and to learn maximum information with the limited data made available to us. To handle the interaction effects and capture non-linearity of the data in a better way, we add a ReLU unit at the end of each fully connected layer. We take the output of these fully connected layers and feed it to the final feed-forward output layer to predict the entity tags for each token.

3.3 Training and Testing

During training, the tokens for each sequence X (generated based on context, definitions and entity types) have annotations of e_k where k is the number of entity types for each token. We calculate categorical cross entropy loss for each token as

$$CE_k = -\sum_{i=1}^{C} t_i log(p_i) \tag{1}$$

where C is the total number of entity classes, t_i is a binary indicator if class label c is a correct prediction for the token k and p_i denote the predicted probability of token k belonging to class c.

The model is trained for 15 epochs with learning rate of $5e^{-5}$ and validated using f1-score on a hold-out sample. We trained the same architecture with various pre-trained transformer architecture such as BERT, RoBERTa and ELECTRA. The models were trained with same architecture so that their performances can be compared.

During inference, the text is passed as tokens and the tags which start with B and continue till I are considered as a single entity and validated accordingly. In practice, we deployed the best performing model among the three architectures to suggest named entities to technicians.

4 Experiments and data

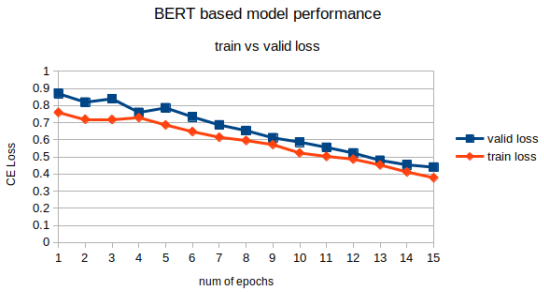


Fig. 3. train and val loss for BERT based NER

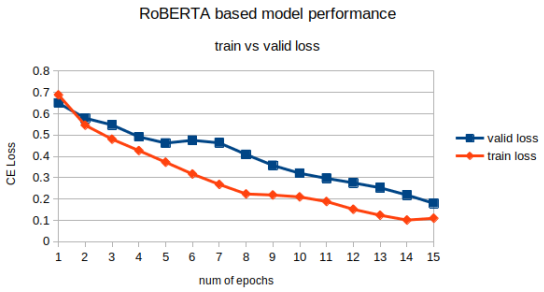


Fig. 4. train and val loss for RoBERTa based NER

4.1 Comparison of performance metrics

The following table 2 shows various performance metrics such as precision, recall and f1-score for the final fine-tuned models based on the various architectures discussed.

We use a learning rate of $5e^{-5}$ for all our experiments. The maximum sequence length was defined as 128 based on what we saw in the training data. This was based on the average sequence length of the technician notes that we encounter. During inference, if we face longer sequence, we break the text into logical chunks of 128 token-long sequences and process that for testing. We use the metrics train loss, validation loss, precision, f1-score for evaluation of model performance.

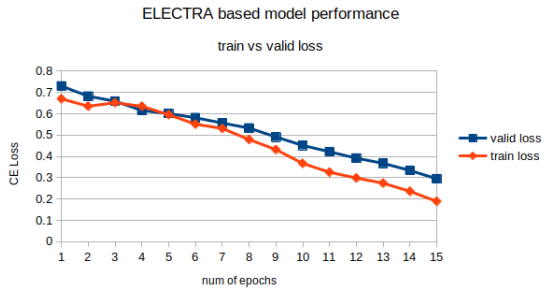


Fig. 5. train and val loss for ELECTRA based NER

	ELECTRA		BERT		RoBERTa	
	TRAIN	VALID	TRAIN	VALID	TRAIN	VALID
LOSS	0.19	0.30	0.38	0.44	0.11	0.18
PRECISION	0.72	0.71	0.65	0.56	0.81	0.80
RECALL	0.77	0.70	0.63	0.56	0.81	0.78
F1-SCORE	0.74	0.69	0.64	0.55	0.81	0.79

Table 2. Overall Performance and loss report

4.2 Comparison of train and validation loss

The train and validation loss for the models fine tuned based on BERT, RoBERTa and ELECTRA based architectures are depicted in 3 to 5 . From the figures we can see that the RoBERTa model has got the least validation and train loss and this is also reflected in the performance of the model for individual entity tags as well.

4.3 Class level performance comparison

The class level metrics for the models were depicted in 3. We compare performance metrics such as train and validation precision, recall and f1. To maintain equal significance to all the classes, including the imbalanced ones, we use macro averaging based F1 measure.

Macro F1 measure is calculated as follows

$$Macro\,F1score = \frac{1}{C}\sum_{i=1}^{N} f1_i \qquad (2)$$

$$where$$

$$f1 = \frac{(2 * precision * recall)}{(precision + recall)}$$

4.4 Inference

The model trained based on the RoBERTa based embeddings performed consistently on all the classes in both train and validation data. This can also be infered

ELECTRA

		Amplitude	Frequency	Replaced Parts	Test error code	model number	Overall
TRAIN	**PRECISION**	0.83	0.64	0.71	0.67	0.81	**0.72**
	RECALL	0.60	0.64	0.79	0.82	0.99	**0.77**
	F1-SCORE	0.69	0.64	0.75	0.74	0.89	**0.74**
VALID	**PRECISION**	0.84	0.58	0.58	0.59	0.83	**0.71**
	RECALL	0.52	0.70	0.69	0.66	1.00	**0.70**
	F1-SCORE	0.64	0.64	0.63	0.62	0.91	**0.69**

BERT

		Amplitude	Frequency	Replaced Parts	Test error code	model number	Overall
TRAIN	**PRECISION**	0.64	0.59	0.70	0.69	0.70	**0.65**
	RECALL	0.48	0.58	0.60	0.61	0.98	**0.63**
	F1-SCORE	0.55	0.58	0.64	0.65	0.82	**0.64**
VALID	**PRECISION**	0.64	0.32	0.57	0.49	0.76	**0.56**
	RECALL	0.49	0.31	0.54	0.42	0.99	**0.56**
	F1-SCORE	0.56	0.31	0.56	0.45	0.86	**0.55**

RoBERTa

		Amplitude	Frequency	Replaced Parts	Test error code	model number	Overall
TRAIN	**PRECISION**	1.00	0.59	0.71	0.76	0.85	0.93	**0.81**
	RECALL	1.00	0.67	0.73	0.75	0.82	0.99	**0.83**
	F1-SCORE	1.00	0.63	0.72	0.76	0.83	0.96	**0.82**
VALID	**PRECISION**	1.00	0.67	0.66	0.74	0.77	0.95	**0.80**
	RECALL	1.00	0.65	0.62	0.69	0.75	0.98	**0.78**
	F1-SCORE	1.00	0.66	0.64	0.71	0.76	0.96	**0.79**

Table 3. Class level performance metrics

from the performance charts and tables. However, the BERT and ELECTRA based models were not able to generalize well on the validation set and hence do not perform as well as the RoBERTa based model. So, for final inference in the product, we used the RoBERTa based model for all the entities.

5 Conclusion and Future Work

For the task of NER, we adopt various pretrained models (BERT, RoBERTa) in this paper. First, we concentrate on data generation utilising unsupervised methods and sequence-to-sequence models. Then, for these entity types, we utilise certain business definitions to eliminate the noisy labels that are generated, and came up with empirical rules iteratively to further minimise the noise in the data. This yields a final annotated dataset that could be utilised in any of the NER training architectures that are based on pre-trained models. Furthermore, we show that RoBERTa based models are better suited to our NER task.

For the purpose of this study, we focused on data quality and quantity and utilized data generation and augmentation techniques to arrive at the best possible training data from unstructured data. In the future, we would be concentrating on few shot learning methodologies that would improve the model's performance. We also want to incorporate active learning based mechanisms to improve model's performance as discussed in [20].

References

1. Paul Thompson and Christopher Dozier. Name searching and information retrieval. 12 2002.
2. Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. *Proceedings of CoNLL-2003*, 03 2004.
3. Martin Hassel. Exploitation of named entities in automatic text summarization for swedish. 2003.
4. Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. Bioelectra:pretrained biomedical text encoder using discriminators. In *BIONLP*, 2021.
5. Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019.
6. Service Intelligence Product, Neuron7.ai. https://www.neuron7.ai/.
7. Hye-Jeong Song, Byeong-Cheol Jo, Chan Park, Jong-Dae Kim, and Yu-Seop Kim. Comparison of named entity recognition methodologies in biomedical documents. *BioMedical Engineering OnLine*, 17, 11 2018.
8. Rafiullah Momand, Shakirullah Waseeb, and Ahmad Latif Rai. A comparative study of dictionary-based and machine learning-based named entity recognition in pashto. pages 96–101, 12 2020.
9. Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. 01 2014.
10. A. Akbik, Tanja Bergmann, Duncan A. J. Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL*, 2019.
11. Sudha Morwal, Nusrat Jahan, and Deepti Chopra. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing*, 1:15–23, 12 2012.
12. Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.
13. Jing li, Aixin Sun, Ray Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 03 2020.
14. Ji Lee, Franck Dernoncourt, and Peter Szolovits. Transfer learning for named-entity recognition with neural networks. 05 2017.
15. Dana Ruiter, Dietrich Klakow, Josef Genabith, and Cristina España-Bonet. Integrating unsupervised data generation into self-supervised neural machine translation for low-resource languages. 07 2021.
16. Omid Jafari, Parth Nagarkar, Bhagwan Thatte, and Carl Ingram. Satellitener: An effective named entity recognition model for the satellite domain. pages 100–107, 01 2020.
17. Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018.
18. Yuwen Zhang. Bert for question answering on squad 2 . 0. 2019.
19. Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
20. G Abinaya, Gyan Ranjan, and P Aswin Karthik. Continuous learning mechanism of nlu-ml models boosted by human feedback. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–6, 2019.

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

AUTHOR INDEX