David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

2nd International Conference on Machine Learning, IOT and Blockchain (MLIOB 2021), August 21~22, 2021, Chennai, India.

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: 978-1-925953-46-6
DOI: 10.5121/csit.2021.111201 - 10.5121/csit.2021.111215

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

Preface

The 2nd International Conference on Machine Learning, IOT and Blockchain (MLIOB 2021), August 21~22, 2021, Chennai, India, 5th International Conference on Signal, Image Processing (SIPO 2021), 5th International Conference on Networks and Communications (NET 2021), 2nd International Conference on Data Mining and NLP (DNLP 2021), 5th International Conference on Software Engineering and Applications (SOEA 2021) and 5th International Conference on Artificial Intelligence, Soft Computing and Applications (AISCA 2021) was collocated with 2nd International Conference on Machine Learning, IOT and Blockchain (MLIOB 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 MLIOB 2021, SIPO 2021, NET 2021, DNLP 2021, SOEA 2021 and AISCA 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, MLIOB 2021, SIPO 2021, NET 2021, DNLP 2021, SOEA 2021 and AISCA 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 MLIOB 2021, SIPO 2021, NET 2021, DNLP 2021, SOEA 2021 and AISCA 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

Abdelhadi Assir, Hassan 1st University, Morocco
Abderrahmane Ez-zahout, Mohammed V University, Morocco
Abhishek Shukla, R D Engineering College, India
AbtoyAnouar, Abdelmalek Essaadi University, Morocco
AdrianOlaru, University Politehnica of Bucharest, Romania
Aishwarya Asesh, Data Scientist II at Adobe, United States
Alex Mathew, Bethany College, West Virginia
Ali A. Amer, Taiz University, Yemen
Alireza Valipour Baboli, University Technical and Vocational, Iran
Amal Azeroual, Mohammed V University, Morocco
Amel Ourici, Badji Mokhtar University of Annaba, Algeria
Ana Luísa Varani Leal, University of Macau, China
Anand Nayyar, Duy Tan University, Viet Nam
Aridj Mohamed, Hassiba Benbouali University Chlef, Algeria
Ashraf Elnagar, College of Computing and Informatics, UAE
Assia Djenouhat, University Badji Mokhtar Annaba, Algeria
BrahimLejdel, University of El-Oued, Algeria
Cheng Siong Chin, Newcastle University, Singapore
Ching-Nung Ynag, National Dong Hwa University, Taiwan
Chin-Ling Chen, Chaoyang University of Technology, Taiwan
Daniel Rosa Canedo, Federal Institute of Goias, Brazil
Dário Ferreira, University of Beira Interior, Portugal
Dariusz Jacek Jakobczak, Koszalin University of Technology, Poland
Dhamyaa Saad Khudhur, Mustansiriyah University, Iraq
Dinesh Reddy, SRM University, India
Domenico Rotondi, FINCONS SpA, Italy
El Habib Nfaoui, Sidi Mohamed Ben Abdellah University, Morocco
ElżbietaMacioszek, Silesian University of Technology, Poland
Endre Pap, Singidunum University, Serbia
Felix J. Garcia Clemente, University of Murcia, Spain
Fernando Zacarias Flores, Universidad Autonoma de Puebla, Mexico
Francesco Zirilli, Sapienza Universita Roma, Italy
Fu Jen, Catholic University, Taiwan
G H Raisoni, College Of Engineering Nagpur, India
GálZoltán, University of Debrecen, Hungary
Gniewko Niedbała, Poznan University of Life Sciences, Poland
Grigorios N. Beligiannis, University of Patras, Greece
Grzegorz Sierpinski, Silesian University of Technology, Poland
Habil Gabor Kiss, Obuda University, Hungary
Hamed Taherdoost, West University, Canada
HamedTaherdoost, University West Canada, Canada
Hamid Ali Abed AL-Asadi, Iraq University College, Iraq
Hamlich Mohamed, UH2C, Morocco
Haouassi Hichem, Abbes Laghrour University Khenchela, Algeria

Hassan Badir,	Abdelmalek Essaâdi University, UAE
Hedayat Omidvar,	Research & Technology Dept, Iran
Hiba Zuhair,	Al-Nahrain University, Iraq
Ikechukwu E. Onyenwe,	Nnamdi Azikiwe University, Nigeria
Israa Shaker Tawfic,	Ministry of Science and Technology, Baghdad- Iraq
J.Naren,	SASTRA Deemed University, India
Jabber,	Vardhaman College of Engineering, Hyderabad, India
Jasmin Cosic,	DB AG, Germany
Jawad K. Ali,	University of Technology, Iraq
Jesuk Ko,	Universidad Mayor de San Andres, Bolivia
JesukKo,	Universidad Mayor De San Andres (Umsa), Bolivia
Jiajun Sun,	Huaiyin Normal University, China
José Luis AbellánMiguel,	Universidad Católica De Murcia, Spain
Jyoti,	Jaypee University, India
Kazuyuki Matsumoto,	Tokushima University, Japan
Kholladi,	Echahid Hamma Lakhdar d'El-Oued, Algeria
Klenilmar Lopes Dias,	Federal Institute of Amapa, Brazil
Larisa Ofelia Filip,	Petroșani University, Romania
LjubomirLazic,	Union University Belgrade, Serbia
LocNguyen,	Independent Scholar, Vietnam
Luisa Maria Arvide Cambra,	University of Almeria, Spain
M V Ramana Murthy,	Osmania University, India
M. ZakariaKurdi,	University of Lynchburg, Virginia, USA
MA. Jabbar,	Vardhaman College of Engg, India
Maad M. Mijwil,	hdad College of Economic Sciences University, Iraq
Mahendra Bhatu Gawali,	Sanjivani College of Engg. Kopargaon, India
Mahsa Mohaghegh,	Auckland University of Technology, New Zealand
Malka N. Halgamuge,	The University of Melbourne, Australia
Maumita Bhattacharya,	Charles Sturt University, Australia
Meenakshi Sharma,	Galgotias University, Greater Noida
Mehdi Gheisari,	Iau, Iran
Mervat Bamiah,	Alnahj for IT Consultancy, Saudi Arabia
Michail Kalogiannakis,	University of Crete, Greece
Mohamed Arezki Mellal,	M'Hamed Bougara University, Algeria
Mohammad Jafarabad,	Iran University of Science & Technology, Iran
Mostafa EL Mallahi,	Sidi Mohamed Ben Abdellah University, Morocco
Mueen Uddin,	Universiti Brunei Darussalam, Brunei
Mu-Song Chen,	Da-Yeh University, Taiwan
Nadia Abd-Alsabour,	Cairo University, Egypt
Nihar Athreyas,	Spero Devices Inc, USA
Noraziah Ahmad,	University Malaysia Pahang, Malaysia
Nour El Houda Golea,	Batna 2 University, Algeria
Noureddin Amaigarou,	Abdelmalek Essaid University, Morocco
Oleksii K. Tyshchenko,	University of Ostrava, Czech Republic
Oliver L. Iliev,	FON University, Republic of Macedonia
Omid Mahdi Ebadati,	Kharazmi University, Tehran
P. S. Hiremath,	KLE Technological University, India
Paulo Batista,	CIDEHUS.UÉ, Portugal
Paulo Jorge dos Mártires Batista,	University of Évora, Portugal
Paulo Trigo,	ISEL/GuIAA, Portugal
Pavel Loskot,	ZJU-UIUC Institute, China
Petra Perner,	Future Lab Artificial Intelligence IBaI-2, Germany

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

2nd International Conference on Machine Learning, IOT and Blockchain (MLIOB 2021)

5th International Conference on Signal, Image Processing (SIPO 2021)

5th International Conference on Networks and Communications (NET 2021)

2nd International Conference on Data Mining and NLP (DNLP 2021)

5th International Conference on Software Engineering and Applications (SOEA 2021)

5th International Conference on Artificial Intelligence, Soft Computing and Applications (AISCA 2021)

IDENTIFYING RANSOMWARE ACTORS IN THE BITCOIN NETWORK

Siddhartha Dalal, Zihe Wang and Siddhanth Sabharwal

Columbia University, New York, USA

ABSTRACT

Due to the pseudo-anonymity of the Bitcoin network, users can hide behind their bitcoin addresses that can be generated in unlimited quantity, on the fly, without any formal links between them. Thus, it is being used for payment transfer by the actors involved in ransomware and other illegal activities. The other activity we consider is related to gambling since gambling is often used for transferring illegal funds. The question addressed here is that given temporally limited graphs of Bitcoin transactions, to what extent can one identify common patterns associated with these fraudulent activities and apply themto find other ransomware actors. The problem is rather complex, given that thousands of addresses can belong to the same actor without any obvious links between them and any common pattern of behavior. The main contribution of this paper is to introduce and apply new algorithms for local clustering and supervised graph machine learning for identifying malicious actors. We show that very local subgraphsof the known such actors are sufficient to differentiate between ransomware, random and gambling actors with 85% prediction accuracy on the test data set.

KEYWORDS

Ransomware Actors Identification, Graph Machine Learning, Local Clustering, Bitcoin Network.

1. INTRODUCTION

Ransomware is a class of malicious software that, when installed on a computer, prevents a user from accessing the computer usually through unbreakable encryption until a ransom is paid to the attacker. In this type of attack, cybercriminals profit from the value victims assign to their locked data and their willingness to pay a fee to regain access to them. Bitcoin is a popular cryptocurrency used by ransomware actors to get ransom as it shields a person's personal identity by allowing them to transact using a Bitcoin address. Further, a bitcoin account holder (i.e., an actor) can create and hide behind multiple bitcoin addresses onthefly. Many fraudulent actors exploit this Bitcoin's pseudo-anonymity for their nefarious purposes. Prominent recent ransomware examples are Locky, SamSam, or WannaCry. As reported by Paquet-Clouston, et al [1], the latter infected up to 300,000 victims in 150 countries and that their lower bound estimate of the amount of bitcoin involved in ransomware transactions between 2013 to 2017 is morethan 22,967.94 bitcoins amounting to over a billion dollarsatthecurrent exchange rate of 1 BTC = $46, 491.11 [02/2021].

The goal of this investigation is to develop systematic ways to identify fraudulent actors in the Bitcoin network through graph classification. This is done by collecting data from multiple public sources on known ransomware addresses reported by their victims. These are used to generate connected transaction graphs in a limited time window. Since an *actor* (i.e., an accountholder) can have many addresses, we identify the bitcoin addresses belonging to the same actor by a new

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
pp. 01-18, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111201

method of local clustering, create features from subgraphs of Actor-to-Transaction bipartite graphs and identify other suspect ransomware actors using supervised machine learning. Figure 1 depicts the overall pipeline. Within the limitations discussed in this paper, we show that we can identify ransomware and gambling actors compared to a random account with accuracy of around 85% on the test dataset.

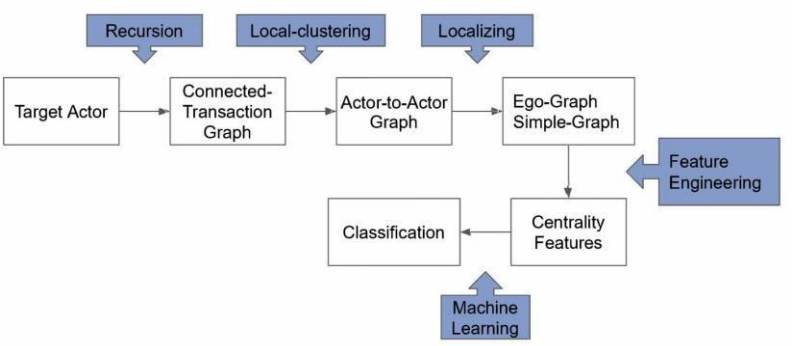


Figure 1. Pipeline of the Approach: Acquisition, creation, wrangling and classification ofdata. Transformations are indicated in blue boxes.

Specifically, Section 2 summarizes some of the previous work. Section 3 discusses the generation of data by web scraping Bitcoin addresses that have been tagged as being a part of a scam by other users in a number of public forums. Section 4 discusses and develops the corresponding temporally limited transaction graphs and the corresponding local clustering strategies. Section 5 proposes ego-graphs generated for ransomware, gambling and random actors for analysis along with several graph centrality metrics as features for supervised learning. Some data analysis is described in Section 6. Strategy for the supervised learning is described in Section 7 with the results of our analysis given in Section 8. Section 9 described limitations of our study and suggests future directions, The final section, Section 10, gives conclusion along with a brief discussion.

2. PREVIOUS WORK

Previous work on this topic can be divided in two parts. The first deals with how to cluster or link various addresses owned by a single actor, and the second discusses ransomware payments.

Many of the so-called behavioral address clustering algorithms are heuristics based. Specifically, Meiklejohn et al. [2] proposed two address-linking heuristics, namely (1) inputs spent to the same transaction are controlled by the same actor and (2) change addresses are not reused. *Change addresses* are commonly used by an account holder to preserve anonymity by creating multiple addresses and transferring bitcoins between those addresses. Indeed, it is considered as a good practice (Nakomoto [3]) to create a new address for transferring the remainder of bitcoins to this newly created address after transferring money to another actor. Harrigan et al [4] highlight that the clustering methods critically depend on the address reuse behavior. Goldfeder et al [5] extends the heuristics to cover CoinJoin transactions. However, Kalodner et al [6]) found that using these set of heuristics resulted in one super cluster of with139 million addresses and many clusters with over 20,000 addresses. This happens mainly because when taking transitive closure of clusters, the errors are propagated across the entire bitcoin blockchain.

Another heuristic approach focuses on tracking IP-addresses, see Biryukov et al [7]. We do not pursue this line of inquiry since the Bitcoin Blockchain doesn't store the IP addresses; it has to be

obtained by getting the log information from e-wallets or mempool. Further, these approaches have a low success rate, from 11% to 60%, as described by Biryukov et al.

While we follow behavioral clustering ideas, we modify them in a number of ways including local clustering. Specifics of algorithms are described in Section 4.

There have also been a number of attempts at using supervised learning to try to classify different categories of actors. For example, Harlev et al [8] uses clustering provided by Chainalysis to classify 434 clusters in different categories that include ransomware, Exchanges, mining pools, gambling, etc. They use also various machine learning algorithms including decision trees, boosting, random forests, etc. They report classification accuracy of 75% and higher. However, since they used clustering provided by Chainalysis, it would not be possible to identify new ransomware addresses without clustering. Further, since their reported results seem to be based on training data with no cross-validation or test data, their results are likely tobe highly optimistic and overfitting (Hastie et al [9]).

Jordan et al [10] also consider a similar problem using graph motifs for classifying Exchanges, Services, Gambling, etc. They mention accuracy of around 90%. The novelty of their approachis the use of graph motifs to derive features for supervised machine learning. However, their analysis doesn't include ransomware actors. Further, their clustering algorithm uses only transitive closure of input addresses. They do not take into account CoinJoin, Coinbase or burn transactions. As mentioned in the clustering part earlier, this is likely to create many false positives. Further, again the accuracy reported seems to be based only on training data. Zola etal [11] also pursue motifs to do supervised learning. They improve results compared to a base model by using cascading classifiers, which uses cascades across 1, 2 and 3 motifs. Their reported results are impressive with accuracy score of up to 98%. However, again they do not consider ransomware actors, nor does their clustering takes into account CoinJoin, change of address or burn transactions.

An alternate approach based on deep learning has been proposed by Jung et al [12]. He proposes the use of Graphical Convolutional Neural Net Models (GCN). Besides being a black box, one of the major problems in using GCN is that each graph for GCN approach hasto have the same number of nodes, which is not the case here. Further, the work did not cluster the addresses, and the reported accuracy of 72% is based on binary classification and only on training data. Our results are better.

Encouraged by Jordan et al [10] and Zola et al [11], we pursue a more detailed approach. Our contribution differ from them in a) our clustering algorithms are local and do take into account CoinJoin, change of address and burn transactions; b) we use ego-graphs instead of motifs - ego-graphs consider relationships between all the actors in the motif graphs, which include triangles, c) our data set includes the ransomware category; d) our features set is basedon various explainable graph centrality metrics; finally e) our analysis is validated by using cross-validation as well as a separate test dataset.

3. CREATION AND WRANGLING OF DATA

3.1. Sources of Data

There were two key sources of data. First, sources that had addresses tagged as being associated with ransomware, i.e., for our "ransomware" class. Second, a source that had a comparison set of addresses that were not associated with ransomware, i.e. for our "random" and "gambling"

classes. The random and gambling addresses are used as a comparison group for supervised machine learning.

Before analyzing the transaction pattern of these addresses, we needed to compile all the Bitcoin transaction data. It was downloaded from the Bitcoin Blockchain using Bitcoin Core [19] and we then accessed the raw data files which contain the validated transactions. Our study includeddata till July 2019 with 400,000,000 transactions and close to 40,000,000 addresses. The binary raw data was converted to a more human accessible format for analysis using BlockSci [6], which is an in-memory analytical database that allows for fast exploration over blocks and transactions due to their sequential, append-only generation process. With this data we had access to the entire transaction history for all addresses.

The first source consisted of creating a database of addresses of known ransomware actors. People who have been victims or approached for ransom, often publish the bitcoin address where bitcoins were asked to be sent as ransomware. Bitcoin WhosWho [21] and Bitcoin Abuse [22] were the two main sites which have such user-submitted addresses. Besides these two sources, we gathered information from the previously published literature [14] and law enforcement published actions (e.g., SEC). Much of the work involved going to these websitesand scrapping the relevant information. All the addresses were collected in 2019.

The second source was Wallet Explorer [20]. It is a website that allows you to view the blocksand the individual transactions inside that block. From there one can also view the addresses and amounts involved in the transaction. The creator of the site Ales Janda wanted to associate addresses with actors. To do this he registered with a variety of businesses that accepted Bitcoinand transacted with those businesses. He then followed the Bitcoins he sent and catalogued which "wallet bitcoins were merged with, or from which wallet it was withdrawn." [20]. He then categorized each one of the businesses he catalogued addresses for into one of five categories: Exchanges, Pools, Services/other, Gambling, and Old/historic. The addresses werecollected from Wallet Explorer in May 2019.

The Old/historic category contains many defunct businesses that early Bitcoin users used. The addresses that are associated with these businesses have been catalogued and the transactions that legitimate users had with these businesses have been web-scraped. Transactions associated with this category are what we call our "Random" class. The reason we call it as such is simplybecause the other businesses Janda compiled were all tagged with a specific category and theseones were across many other miscellaneous categories.

The gambling addresses we have collected are all of the Bitcoin addresses that have sent or received money from any of the associated gambling websites like CoinGaming, PocketDice, and BitcoinPokerTables but are not directly tagged to those websites. These websites are primarily designed so users can gamble using Bitcoin, however they sometimes have the addedconsequence of allowing money laundering to occur as users can "clean" their stolen Bitcoin into cash [13]. Transactions associated with these gambling addresses are what we call our "Gambling" class.
In total we collected 143, 498 and 216 ransomware, random and gambling addresses.

3.2. Graph Creation

For analysis, we need to extract relevant data of ransomware, gambling and random addresses from the Bitcoin Core. Bitcoin has three primary connected components: transactions, addressesand bitcoin transferred. From these, transactions-address bipartite graph can be created. Transactions are arranged in a set of sequentially linked blocks generated randomly approximately every 10 minutes (around 144 blocks per day). Each transaction has a set of input addresses, a set of output

addresses and the amount of bitcoins transferred between them. There is also a transaction fee paid to the miner- an address that created the block.

However, there is no input for the Coinbase transactions, which are algorithmic transfers of bitcoins, one in each block to the corresponding miner.

3.3. Transaction Graphs

In the simplest terms, the transaction graph consists of a directed acyclic graph (dag), ***T, A,W***, with transactions in the set ***T*** as nodes, input to output addresses in the set ***A*** as directed edges between transactions, and the bitcoin transferred in the set ***W*** as edge weights. Except for the transactions in the first block (the so-called genesis block) and Coinbase transactions, each transaction node is connected to multiple previous transactions as input nodes. A transaction node may not have an output node at a given time since there may be no transactions utilizing the output addresses of that transaction as input address so far. Specifically, a directed edge from node X to node Y means that an output address in X was an input address in Y and spent all of the Bitcoin they received in X in Y.

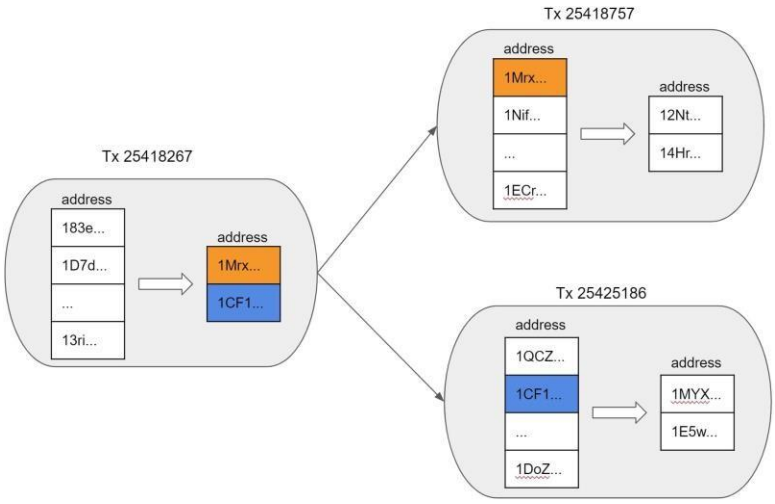


Figure 2. Anatomy of Transaction Graph: the transaction on leftis connected by two different addresses (orange and blue) to twodifferent transactions on right

Obviously, the transaction graph of the bitcoin transaction starts from the genesis block to the last block being considered. However, our interest lies in the behavior of the entity represented by a given address with the hope of identifying common patterns. Thus, we look at only the transactions involving an address under consideration (a random, gambling or ransomware address) and extend it iteratively to the transactions feeding the actor's transaction and the transactions being fed by the actor's transactions. Towards that end, given an actor *A*, we identify the first transaction T_A involving that actor, and iteratively identify the transactions T_p feeding to T_A and transactions T_f being fed by it.

This defines children-parents relationship. This process is followed iteratively taking transitive closure of all the children and parent transactions of T_p and T_f in the set of all transactions T. Though this limits the cardinality of the newly formed transaction set $\boldsymbol{T_A}$ corresponding to an address *A*, the beginning of the chain can still reach the genesis block or Coinbase transactions.

$\boldsymbol{T_A}$ can be further reduced for our problem. Based on looking at the behavior of a number of

ransomware actors in their corresponding transaction graph $\boldsymbol{T_A}$, it was felt that just like many other criminals, the ransomware artist, after receiving the payment would rapidly make a succession of transactions to other addresses to make tracing difficult. Thus, it was decided to concentrate only on temporally local behavior of the address in $\pm n$ blocks. Specifically, for defining this local behavior, we restrict T_A further from our recursion by using the following 3 rules.

1. Given a first transaction T_A *in* $\boldsymbol{T_A}$ by the address A under consideration, let the set $\boldsymbol{T_{A,n}}$ in $\boldsymbol{T_A}$ represent all the transactions in blocks $\pm n$ height away from the block containing the transaction T_A. For example, if transaction T_A was in block $10,000$ then the set $\boldsymbol{T_{A,n}}$ represents all the transactions between blocks $10,000 \pm n$.
2. We further restrict $\boldsymbol{T_{A,n}}$ when the output side is an address belonging to an exchange or gambling business as identified by Wallet Explorer. We do this because the children transactions links of the exchange node have many actors that have nothing to do with each other. The analogy is with a bank or a casino; if an actor was to deposit money in the bank or buy chips at the casino, we do not want to follow all the other actors who dealt with the bank or redeemed chips at the casino as they aren't necessarily linked to the actor of interest.
3. As an exception to the stopping criterion 1, we do not follow a Coinbase transaction backward, where a miner is awarded new bitcoins since there are no parent transactions.

Besides these criteria, we restricted $\boldsymbol{T_{A,n}}$ further as stated below:

- *Non-standard Scripts*. There were several cases that Blocksci or various other explorers could not parse the address and would return a NaN, which can result in a situation that the output of source transaction or the input of the destination transaction or both are NaN. In this situation, in order to prevent the loss of information, we created dummy addresses to replace the NaN, unless we found an explorer which could parse the script. In that case we manually inserted the correct address.
- *Proof of Burn. OpReturn* transactions where an address may burn bitcoin to save a data item on the blockchain were assigned a string 'burn' to replace the NaN.
- *CoinJoin Transactions.* For the Actor-to-Actor graph creation, we need to identify mixing CoinJoin transactions. For more details on CoinJoin and other mixing transactions we refer to [15]. Given the new services like *Wasabi* [23] and *Samorai* [24], the older proposed identification rules do not work. We empirically modified rules used by BlockSci to tag CoinJoin transactions with the following rules.

  - If the transaction has less than 2 input or 3 output addresses, it is not a CoinJoin.
  - if the number of input addresses is smaller than the half of the number of the output addresses, the transaction is not a CoinJoin.
  - if the number of the output addresses is less than 6 and all output amounts are equal, the transaction is considered a CoinJoin.
  - if the number of the output addresses is more than 6, the transaction is considered CoinJoin if at least 5 output amounts are equal.

In summary, $\boldsymbol{T_{A,n}}$ is the connected subgraph of the first transaction involving the actor A within n blocks either side with the exceptions described in the previous paragraphs.

For building the weighted transaction graphs, each of the edge of transaction graph had weight corresponding to transacted bitcoins. This required the information on the amount of input and output bitcoins and transaction fees in each transaction, maintaining the equilibrium - namely:

Input_amount = output_amount + transaction fees

As an example of $T_{A,n}$ Figure 3 depicts a sample of the directed transaction graph emanating from the Actor ”12*HaVrpXkLr*2*UnkM f* 6*X* 9*b* on both sides. For ease of viewing all the self-loops have been removed, multiple edges have been collapsed and with no weights.

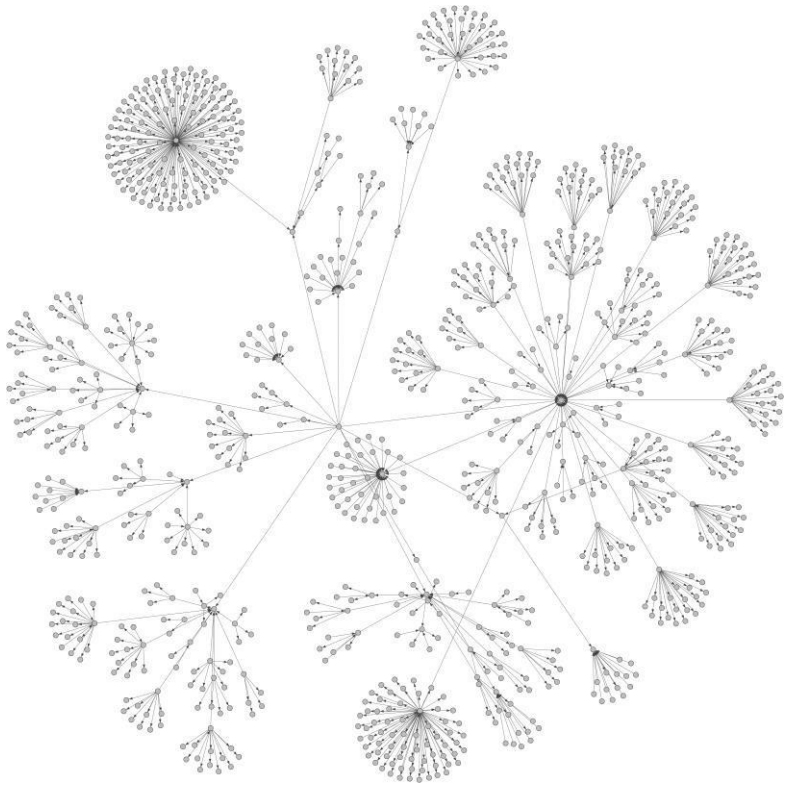


Figure 3. Actor 12HaVrpXkLr2UnkMf6X9bY11cuNrZUdUnV Transactions

4. CREATION OF LOCAL CLUSTERING AND ACTOR-TO-ACTOR GRAPHS

Since we are interested in the behavior of an actor, we need the corresponding actor-to-actor graph (not address-to-address graph). There are several difficulties with this. The main one lies with identification of the set of all addresses used by an actor. This is mainly, as mentioned earlier, due to bitcoin network allowing an account holder to create multiple bitcoin addresses on the fly. For simplicity we call the set of all addresses owned by an actor as an *entity* set and the corresponding graph is called an *entity graph* or an *Actor-to-Actor graph*.

As it is widely accepted, any address clustering scheme is imperfect, and ground truth is difficult to obtain on a large scale, since it requires interacting with service providers. Many other heuristics are possible, including those that account for the behavior of specific wallets. In our experiment with the data going to July 2019, when applying such heuristics to the entire blockchain, we got one super cluster containing more than 90% of addresses. This is primarily due to tumblers (the services which mix bitcoins) and CoinJoin kinds of transactions where multiple parties combine their transactions to preserve their anonymity. This is compounded by misattribution of the change of address.

We considered several modifications to the basic logic of behavioral clustering. But, when applied globally, all of them have exceptions which make a large number of false unions resulting

in large clusters due to transitive closures. To limit potential for wrong clusters which gets propagated across the entire bitcoin blockchain, we adopt a different strategy of creating local clusters since our objective is mainly to identify scam artists who try to move bitcoin in a short period of time soon after starting their ransomware related scam. As discussed by Kharraz et al [16], just like any other crimes, ransomware artists move ransomware payments as quickly as possible. Thus, we decided to apply the clustering algorithm discussed earlier only locally within the temporal limit of n=±144 blocks; basically within ±1 day.

We also apply somewhat different logic to CoinJoin and other transactions as described in the previous section. We may still have some false positives, but we will never have super clusters due to transitive closure. Any large clusters effects will be felt only within a particular local graph and not globally.

Specifically, we use the following rules:

1. Inputs spent to the same transaction are controlled by the same actor, thus, the entity set is the union of all those addresses.
2. If there is only one change address, identified by it never being used prior to the current transaction, consider it as a part of the input address set.
3. Exceptions to the rules 1 and 2 is when a transaction is identified as CoinJoin. In that case do not take a union.

The pseudo-code for the clustering process is given below.

Algorithm 1 Generate Local Cluster

for all transactions in the graph **do**
if this is a CoinJoin transaction **then**
assign each input and output addresses as separate clusters.
else
if there is only one new address in output addresses of this transaction assign all input addresses and output addresses as one cluster.
else
assign all input addresses as one cluster;
assign each input and output addresses as separate clusters.
end if **end if**
end for
After collected all address-cluster mapping, merge the mapping till there is only one cluster for each address

For the weighted graph analysis, we need to find the bitcoin transfer between addresses. Since the bitcoin transfer is defined between the sets of input addresses and output addresses, there is no exact way to allocate the amount between a given input address and an output address unless one of the sets has cardinality 1. We approximate the transfer by a proportional allocation rule. Namely, given a transaction with input addresses: $I_1, \ldots, I_x$; input amounts: $IA_1, \ldots, Ia_x$; output addresses: $O_1, \ldots, O_y$; and output amounts: $OA_1, \ldots, Oa_y$, the edge weight from I_i to O_j is computed by the following formula: $(Ia_i/\Sigma IA_k)*Oa_j$, which is further adjusted by the transaction fee. This is an approximation. For Actor-to-Actor graph, the weights are the sum of all individual weights of the corresponding addresses.

The development described so far allow us to form Actor-to-Actor weighted graphs. Some of these graphs after clustering had only a small number of nodes and were deleted. These were

further sub-divided in training and test sets of 328 and 82 graphs (80−20%), respectively, for supervised learning.

5. GRAPHS AND CENTRALITY FEATURES

5.1. Subgraphs

Recall that the following the logic of the last sections, we consider the locally clustered *Actor-to-Actor* graph from the connected transaction graph $\boldsymbol{T_{A,1}}$ for an actor *A* within ±144 blocks (±1 day). The subgraph of all addresses within $\boldsymbol{T_{A,1}}$ is referred to simply as *whole* graph. For additional analysis, we take several different kinds of subgraphs. Specifically, since our primary interest lies in the actor under consideration, we created Ego subgraphs for the actor. *Ego graph* of order n of a node is the subgraph formed by the nodes that are within the neighborhood of order *n* of the node without considering the direction of the edges. Ego graphs are richer than standard motifs since they also consider relationships between neighbors. Further, another set of subgraphs, called *simple graphs* were obtained by removing loops of the nodes to itself, and collapsing multiple edges to one edge. These subgraphs are considered since it is expected that the actor's footprints would be most visible in its direct transaction with other nearby actors. For example, the ransomware actor's footprints would be most visible in its interactions with the victims and other nearby actors and co-conspirators. For further analysis we only considered ego1, ego2, ego3 and the corresponding versions of simple graphs.

5.2. Centrality features

For each of the subgraphs discussed above, we extracted a number of graph-based features:

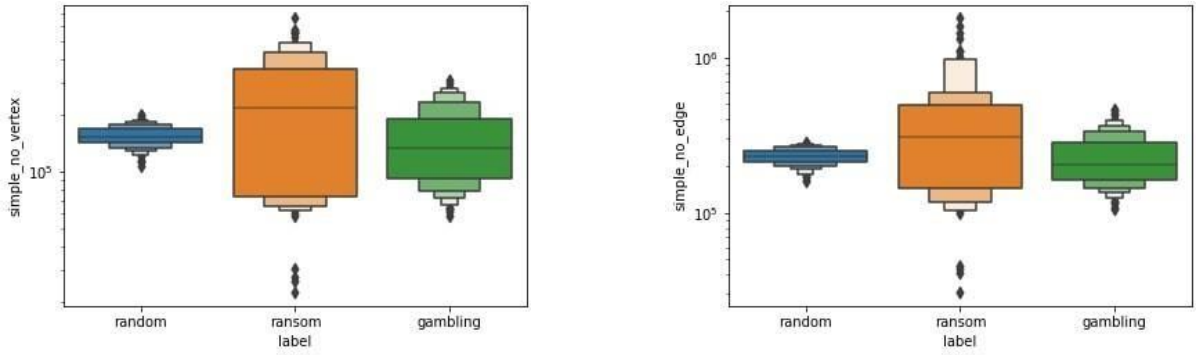
i. Basic Statistics: # of Vertexes, # of Edges, Total bitcoins, Loops, Degree, Neighborhoodsize
ii. Centralities: Normalized Closeness, Betweenness, Page Rank, Cluster Coefficient, Coreness, Hub and Authority

Definitions of each can be found in **igraph** [26] with more details in article [17]. Some are overall graph parameters; the rest are restricted to the node of the actor under consideration. A number of variants of these were considered where it made sense including weighted, unweighted and directed. The creation of graph and extraction were all carried out by using *Python igraph* library [26].

The task of computing graphs and its features was computationally intensive. For efficiency reasons, we did not consider the graphs larger than one million unique addresses or more than 1/2 million transactions. As mentioned earlier, *whole* graph with only a small number of nodesand the corresponding ego graphs were also removed, Finally, to balance the classes better a random sample of size 155 was taken from random graphs. An *80-20 split* between training and test data with stratification resulted in the training set of 328 *whole* graphs (124 random, 80 ransom, 124 gambling), and the test set of 82 *whole* graphs (31 random, 20 ransom, 31 gambling).

6. EXPLORATORY DATA ANALYSIS

In this section, we describe an exploratory analysis on some of the features generated in section 5. Here, we used the 'boxen plot' which centers a distribution at its median line; each successive level outward contains half of the remaining data until it reaches to the outlier level. For details, see *seaborn* [27].



(a) # of vertices (b) # of edges

Figure 4. whole-simple graph

Figure 4 shows the number of vertices and edges in the whole-simple Actor-to-Actor graphs. *Recall that the whole graph is based on recursion of all connected transactions associated within 2 days of the actor*. Thus, these graphs could be skinnier than the graphs over two days dependingupon the level of connections of the actor. We can see that for 'random' and 'gambling' graphs, the distribution does not differ a lot. However, there are many extreme values in 'ransom' graphs and it is flatter compared to other two. It reflects the nature of the 'ransomware' class where actors will try to obfuscate their transaction patterns through complicated laundering, which also reflects that the local clustering algorithm performs well.

For brevity, the rest of the analysis highlights only Ego-1-simple graphs for a few important features found in the Results section since they are the simplest graphs of other actors who arein close touch with the Actor. The analysis looks at the marginal distribution of the selected features across all the actors separated by ransomware, random and gambling categories.

PageRank, also known as Google Rank, is a way of measuring the importance of website pages.The assumption is that more important websites are likely to receive more links from other websites. For more definition, see *PageRank* [28]. As seen in Figure 5, the 'ransom' clusters tend to have a higher PageRank, which means it is likely to receive more transactions from other clusters. This makes senses when the ransomware attack is one wave after another and usually sent to lots of users within a short period of time. Also, PageRank of random actor on average seems to be lower than ransomware actors indicating that the ransomware actors are more often recipients of funds.

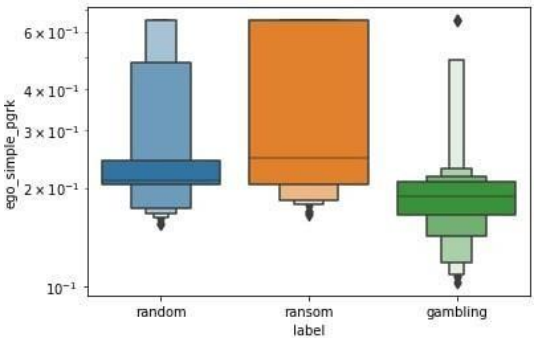


Figure 5. Page-Rank

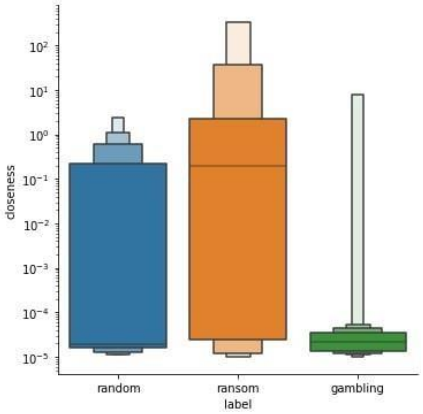


Figure 6. Closeness

The closeness centrality of a vertex measures how easily other vertices can be reached from it (or the other way: how easily it can be reached from the other vertices). For definition, see *closeness* [26], [31]. The weighted-IN closeness of ego-1 simple graphs is shown in Figure 6. The 'gambling' tends to have less centrality, which shows similar pattern as shown in Figure 5. This suggests that gambling actors are not closely connected to other accounts. They have many more outliers indicating that there are few very large gamblers and possibly the distribution is scale-free.

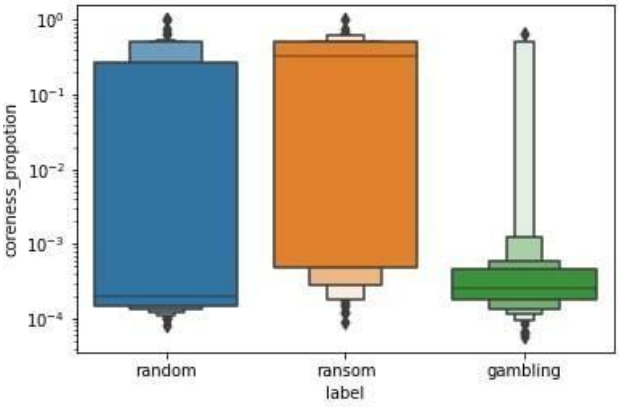


Figure 7. Coreness normalized by # of vertices

In Figure 7, we show the *coreness*(All). The k-core of graph is a maximal subgraph in which each vertex has at least degree k. The coreness of an Actor is k if it belongs to the k-core but not to the (k+1)-core. For the definition, see *coreness* [29]. The coreness across graph is normalized by the number of vertices since all the graphs have different number of vertices. From the figure, as before, it can be noticed that the gambling graphs have a relatively low coreness.

Finally, Figure 8 shows, as one would have expected that cluster coefficients for gambling class is much smaller and followed by ransomware class since both of those classes are directly involved in possibly criminal activities and they would minimize their interactions with actors which are more connected with each other.

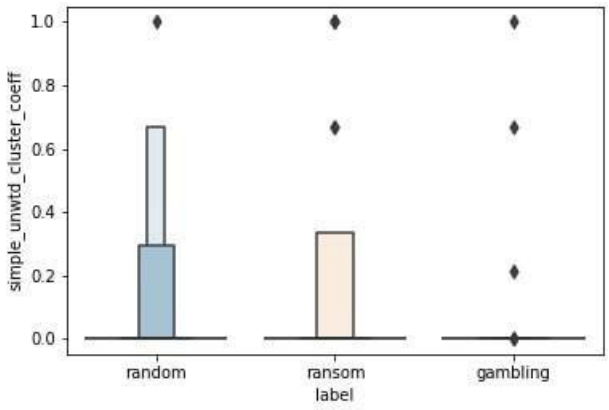


Figure 8. Unweighted Cluster Coefficient, Normal Scale

The comparative analysis of the marginal distributions of features reported so far suggests that different classes behave somewhat differently from each other. For example, gambling actors behavior is rather different than other actors in closeness, PageRank, cluster-coefficient and coreness. Further, the PageRank of the ransomware actors is higher. This analysis indicates that these features could be good candidates for any machine learning model.

7. MACHINE LEARNING ON ACTOR-ACTOR GRAPHS

For the purposes of supervised learning, the extracted whole graphs were divided in testing (20%) and training (80%) graphs stratified by their categories. Further, for each whole graph only the subgraphs of ego-graph 1, ego-graph 2, ego-graph 3 and their simple counterparts were extracted for analysis because of their proximity to the actor under consideration. Only subsets of features given in Table 1 were extracted from each of these graphs. The subset was obtained by keeping only one of each set of highly correlated features. The graphs and the corresponding number of features are shown in Table 1 below.

Table 1: Centrality Features Considered

	ego3	ego3-simple	ego2	ego2-simple	ego1	ego1-simple
# of features	11	16	13	16	12	11

7.1. Models

We consider supervised learning in three stages shown in Table 2.

Table 2: Modeling Strategy/Stages

Learning	Initial	Intermediate	Final
Type	Multiple Classifiers	Stacking	Bagging

In the Initial stage multiple classifiers were fitted to each of the 6 sub-graphs. Since each classifier has different strengths and weaknesses in different regions of the feature space, as an intermediate model, we used ensemble learning technique of Stacking [18] to improve. Specifically, the stacked model uses the predicted probabilities of each class by each classifier as features and predicts the probability of each class by using a simple model (logistic in our case). Even though we have 3 classes, since the probabilities add up to 1 for each of the six classifiers,

there are 12 such linearly independent features. This process is depicted in Figure 9 with the six different base classifiers we used for creating the ensemble.

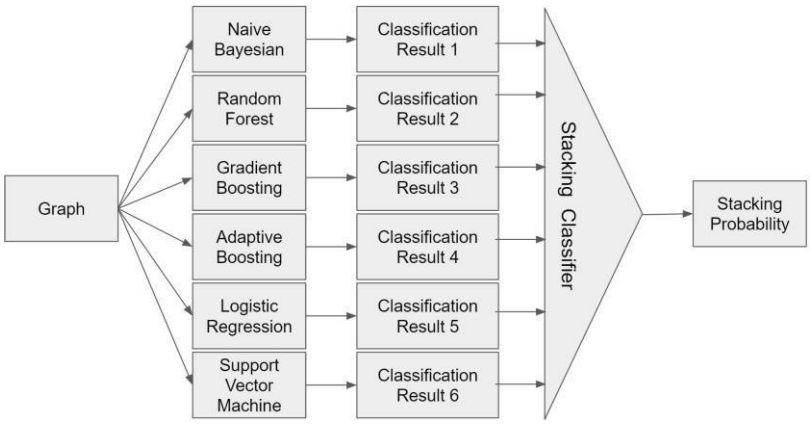


Figure 9. Stacking Model

In the final Stacking-Bagging stage, we combine the results across different subgraphs by creating a meta classifier (or called Final Classifier). It is a simple classifier (in our case logistic)that uses the probabilities of each class in the subgraph stacked models as the feature set and fuses them. This is analogous to bagging since we have 6 different data sets (subgraphs) each containing estimated probabilities of each class. Just like in stacking we will have 12 features for the six types of sub-graphs. The final attribution of the class is given to the class with highest probability by the meta-classifier. This process is depicted in Figure 10. We used the predicted probabilities from 6 graphs as new features and trained a final meta-classifier. We called this as 'stacking-bagging' model.

To motivate efficacy of the stacking-bagging model, consider the simple fusing by averaging the probabilities across 6 estimated probabilities for each data set. In that case, Mean Squared Error (MSE) = Bias^2 +Variance. Each component on the bias term is roughly the same constant since they are using the same type of estimators. The second term=average_variance/6+∑Covariances/6. In our case, since each estimated probability use different graphs, it is expected that the covariances would be relatively negligible. Thus, the MSE will be substantially less compared to non-fusing.

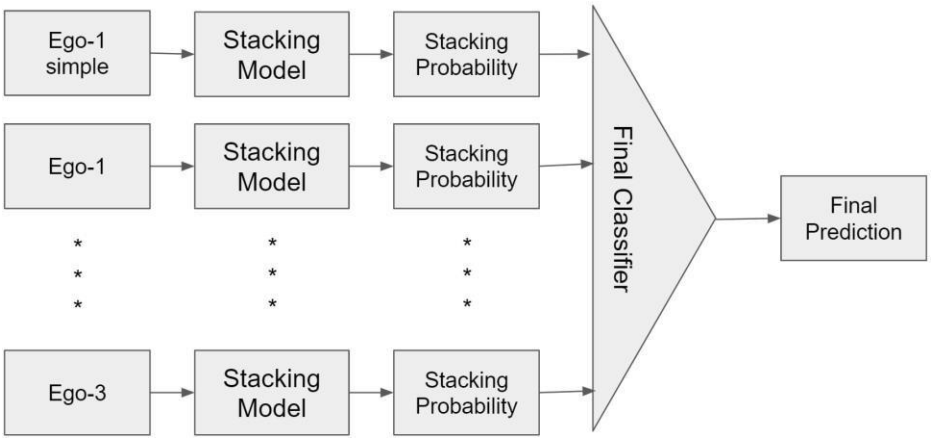


Figure 10. Stacking-Bagging Model

We used the cross-validation score on balanced-accuracy as our objective and finally ran our model on our test set.

7.2. Cross Validation and Efficacy Metrics

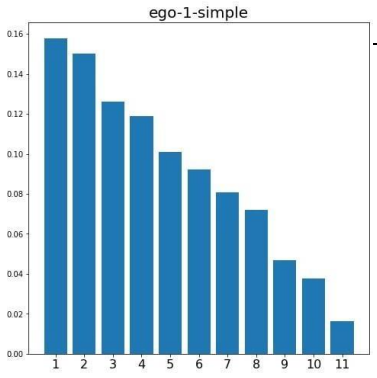
To implement the strategy outlined in 6.1, and to measure its efficacy we used the training data with cross validation for model selection. Specifically, when training the classifiers with grid-search and cross-validation, we used 5-folds with stratification on labels and 80% of the data for train-validation and 20% for testing. For the meta-classifier in the stacking model and for the stacking-bagging model, we used Logistic Regression.

Since we have a multi-label classification problem, we used balanced accuracy or simply Accuracy in tables), weighted precision (as Precision in tables), and weighted recall (as Recall in tables) as our evaluating metrics. We refer to them simply as accuracy, precision and recall. For definitions see *scikit-learn* [25].

8. RESULTS

Table 3 gives cross-validated accuracy on training data of the 6 base classifiers using the cross-validation on the training data. Gradient boosting and Random Forest models seem to outperform others with around 75% to 80% balanced accuracy. Taking a deep dive on feature importance, Figure on the left side of Table 3 shows the 11 features included in Random Forest model for the ego-1 simple graph. Besides the commonly used graph centrality features, it includes less used features like *coreness* and *cluster coefficient*.

Table 3: Feature Importance of Ego-1-simple graph



Rank	Feature	Value
1	Closeness(wtd/out)	0.158
2	sum of weights	0.150
3	Closeness(uwtd/out)	0.126
4	# of vertices	0.119
5	Closeness(wtd/in)	0.101
6	cluster coefficient	0.092
7	Closeness(wtd/all)	0.081
8	Closeness(uwtd/in)	0.072
9	Coreness(all)	0.047
10	Authority	0.038
11	Coreness(IN)	0.016

Table 4: Balanced Accuracy of Classifiers for each graph

	ego-1-s	ego-1	ego-2-s	ego-2	ego-3-s	ego-3
Naive Bayesian	0.4148	0.4324	0.5514	0.4034	0.5189	0.4047
Random Forest	0.7346	0.7924	0.7596	0.7966	0.7973	0.8171
Gradient Boosting	0.7255	0.7709	0.7456	0.7651	0.7829	0.8107
Adaptive Boosting	0.6911	0.6937	0.7111	0.7160	0.7444	0.7386
Logistic Regression	0.6162	0.6272	0.5862	0.6000	0.6101	0.5962
SVM	0.6126	0.6271	0.5997	0.6031	0.6298	0.5908

Stacking these models produces cross-validated balanced accuracy between 96% and 99%, a

substantial improvement. It is interesting to note that ego-simple graphs tend to outperform their corresponding ego graphs.

Table 5: Stacking Model: Performance by Cross-Validation

	Accuracy	Precision	Recall
Ego-1-simple	0.9932	0.9942	0.9939
Ego-1	0.9890	0.9910	0.9909
Ego-2-simple	0.9602	0.9661	0.9634
Ego-2	0.9662	0.9690	0.9664
Ego-3-simple	0.9917	0.9941	0.9939
Ego-3	0.9743	0.9735	0.9724

In the final stage, we build a bagging-stacking model on all six types of ego subgraphs leading to the cross-validated accuracy of 1 and 85% on the test set as shown in Table 6. Thecorresponding confusion matrix is given in Figure 11.

Table 6: Final Model Accuracy, Precision and Recall

	Cross-Validation	Test
Accuracy	1	0.8537
Precision	1	0.8566
Recall	1	0.8537

As seen from the above table, the final model outperformed the stacking model as measured by cross-validated accuracy of 1. Figure 11 gives the corresponding confusion matrix for the final model on the test set. There is no systematic confusion evident in the confusion matrix.

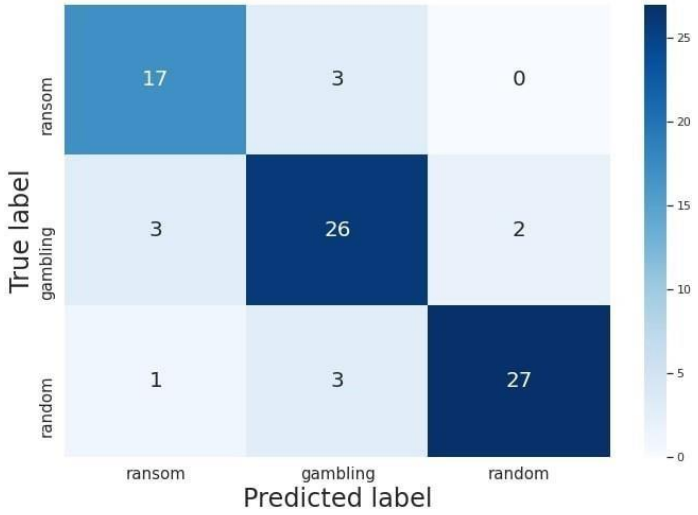


Figure 11. Confusion Matrix of the test data

9. LIMITATIONS AND FUTURE WORK

Though our results indicate high accuracy, further improvement should be possible by improving clustering algorithms, better ways to associate actors involved with non-standard scripts and

Coinjoin transactions. Further, our data set is limited consisting of around 400 graphs, each with thousands of nodes. Further, our ground truth is based on what is user- reported. Getting more data which is more reliable would improve accuracy. Also, given that we are using stacking, it is hard to interpret the final model. More interpretative models using a different machine learning approach may be feasible. Finally, though we have not undertaken it here, it would be worthwhile trying to identify actors from TOR since they are also likely to be involved in illegal activities [30]. Another direction to explore would be to see how these techniques can be generalized to other alternative crypto-currencies.

10. CONCLUSIONS

This paper addresses the key question of how to identify miscreants who are involved in ransomware and in gambling compared to random actors. The problem is difficult due to the pseudo-anonymity of the Bitcoin network. Specifically, the question addressed here is that given temporally limited graphs of Bitcoin transactions, to what extent can one identify common patterns associated with these fraudulent activities and apply them to find other similar actors. The singular contributions of this paper include a) extraction and creation of transaction graphs associated with the miscreant actors, b) clustering all the nodes of such graphs in common entities controlling those accounts while taking into account different kinds of transactions, c) using supervised learning novel algorithms to create models based on actor to actor ego graphs that identify similar miscreants, d) validating the models on cross validated data with accuracy of 1 and on the test data set of around 85%.

11. ACKNOWLEDGEMENTS

We acknowledge Professor Lazaros Gallos, Jianqiong Zhan, Vatsal Randhar and Xiaoqi Wang for helpful discussions and contributions. Financial support from School of Professional Studies, Data Science Institute and Statistics Department at Columbia University is also gratefully acknowledged. Computing support was provided by Habanero High Performance Computing Cluster at Columbia University

REFERENCES

[1] Paquet-Clouston M, Haslhofer, B et al. “Ransomware payments in the Bitcoin ecosystem”, *Journal of Cybersecurity* (2019).

[2] Meiklejohn, S. et al. “A fistful of bitcoins: characterizing payments among men with no names.” *Proceedings of the 2013 conference on Internet measurement conference* (2013).

[3] Nakamoto, Satoshi. “Bitcoin: A Peer-to-Peer Electronic Cash System.” (2009).

[4] Harrigan, M. and Christoph Fretter. “The Unreasonable Effectiveness of Address Clustering.” *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress(UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (2016): 368-373.

[5] Goldfeder, Steven et al. “When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies.” *Proceedings on Privacy Enhancing Technologies* 2018 (2018): 179 - 199.

[6] Kalodner, Harry A. et al. “BlockSci: Design and applications of a blockchain analysis platform.” *ArXiv* abs/1709.02489 (2020).

[7] Biryukov, A. et al. “Deanonymisation of Clients in Bitcoin P2P Network.” *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014)

[8] Harlev, Mikkel Alexander et al. “Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning.” *HICSS* (2018).

[9] Hastie, T., Tibshirani, R., Friedman, J. (2008), The Elements of Statistical Learning, Springer

[10] Jourdan, M. et al. “Characterizing Entities in the Bitcoin Blockchain.” *2018 IEEE International*

Conference on Data Mining Workshops (ICDMW) (2018): 55-62.
[11] Zola, Francesco et al. "Cascading Machine Learning to Attack Bitcoin Anonymity." *2019 IEEE International Conference on Blockchain (Blockchain)* (2019): 10-17.
[12] Jung, K. "Bitcoin Ransomware Detection with Scalable Graph Machine Learning", *YOW! Data Conference* (2019).
[13] Fanusie, Y and Robinson, T. "Bitcoin Laundering: An Analysis of Illicit Flows into Digital Currency Services," *Foundation for Defense of Democracies* (2018), http://defenddemocracy.org/content/uploads/documents/MEMO_Bitcoin _Laundering.pdf.
[14] Conti, M. et al. "On the Economic Significance of Ransomware Campaigns: A Bitcoin Transactions Perspective." *Comput. Secur.* 79 (2018): 162-189.
[15] Camino, R. et al. "Finding Suspicious Activities in Financial Transactions and Distributed Ledgers." *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (2017): 787-796.
[16] Kharraz, Amin et al. "Cutting the Gordian Knot: A Look Under the Hood of Ransomware Attacks." *DIMVA* (2015).
[17] Saxena, Akrati and S. Iyengar. "Centrality Measures in Complex Networks: A Survey." *ArXiv* abs/2011.07190 (2020).
[18] Wolpert, D.. "Stacked generalization." *Neural Networks* 5 (1992): 241-259.
[19] "Download Bitcoin Core." Bitcoin, bitcoin.org/en/download.
[20] *WalletExplorer.com: Smart Bitcoin Block Explorer*, www.walletexplorer.com/.
[21] *Bitcoin Address Lookup*, https://bitcoinwhoswho.com/
[22] *Bitcoin Abuse Database*, www.bitcoinabuse.com/.
[23] *Wasabi Wallet - Bitcoin Privacy Wallet with Built-in CoinJoin*, www.wasabiwallet.io/.
[24] *Samourai Wallet*, samouraiwallet.com/.
[25] "3.3. Metrics and Scoring: Quantifying the Quality of Predictions." *Scikit*, scikit-learn.org/stable/modules/model_evaluation.html# balanced-accuracy-score.
[26] "The Network Analysis Package." *Igraph*, igraph.org/.
[27] "Seaborn Boxenplot." *Seaborn.boxenplot - Seaborn 0.11.1 Documentation*, seaborn.pydata.org/generated/seaborn.boxenplot.html#seaborn.boxenplot.
[28] "PageRank." *Wikipedia, Wikimedia Foundation*, 20 Mar. 2021,en.wikipedia.org/wiki/PageRank.
[29] Batagelj, V. and Matjaz Zaversnik. "An O(m) Algorithm for Cores Decomposition of Networks." *ArXiv* cs.DS/0310049 (2003)
[30] Nabki, Mhd Wesam Al et al. "Classifying Illegal Activities on Tor Network Based on Web Textual Contents." *EACL* (2017).
[31] Freeman, L.C. "Centrality in Social Networks I: Conceptual Clarification." *Social Networks* (1979), 1, 215-239.

AUTHORS

Siddhartha Dalal is Professor of Practice at Columbia University. He received his MBA and PhD from University of Rochester. Prior to joining Columbia, he was the Chief Data Scientist and Senior VP at AIG, CTO at RAND Corporation, VP of Research at Xerox and Chief Scientist and Executive Director at BellLabs/Bellcore. He also advised the US Army and DoD on technologies. He has over 100 peer-reviewed publications, patents, and monographs covering the areas of risk analysis, medical informatics, Bayesian statistics and economics, image processing, and sensor networks. He has received several awards including from IEEE, ASA, and ASQ, notably for his work on Space Shuttle Challenger disaster and for managing software risks. The US Army has awarded him the Meritorious Civilian Service Medal.

Zihe Wang is currently a research staff associate at Columbia University. He completed his Master's in data science from Columbia University in 2020 and his Bachelors in Statistics and Computer Science from University of Illinois at Urbana-Champaign in 2019.

Siddhanth Sabharwal is currently a first-year PhD student in the Statistics department at University of Illinois at Urbana-Champaign. He completed his Master's in Statistics from Columbia University in 2019 and his Bachelors in Statistics and Computer Science from University of California Davis in 2017.

SUMMARIZATION OF COMMERCIAL CONTRACTS

Keshav Balachandar[1], Anam Saatvik Reddy[1],
A. Shahina[1], Nayeemulla Khan[2]

[1]Department of Information Technology,
SSN College of Engineering, Chennai, India
[2]School of Computer Science and Engineering, VIT University, Chennai, India

ABSTRACT

In this paper, we propose a novel system for providing summaries for commercial contracts such as Non- Disclosure Agreements (NDAs), employment agreements, etc. to enable those reviewing the contract to spend less time on such reviews and improve understanding as well. Since it is observed that a majority of such commercial documents are paragraphed and contain headings/topics followed by their respective content along with their context, we extract those topics and summarize them as per the user's need. In this paper, we propose that summarizing such paragraphs/topics as per requirements is a more viable approach than summarizing the whole document. We use extractive summarization approaches for this task and compare their performance with human-written summaries. We conclude that the results of extractive techniques are satisfactory and could be improved with a large corpus of data and supervised abstractive summarization methods.

KEYWORDS

Text summarization, automatic summarization, commercial contracts.

1. INTRODUCTION

In today's day and age, contracts are drafted for every agreement between two parties, documents that companies, firms, and individuals deal with are increasing rapidly. It has become very difficult for corporate staff and chief officers to review contracts which could either be 2 pages or go beyond 100s of pages. To alleviate this difficulty, a large number of companies engage tools for summarizing contracts, extracting key pieces of information, and aiding in other such tasks. Summarization of the entire document is not fruitful as the summaries might be too vague and each line carries a different level of importance. This has been the main motivation behind our project. Thus we propose a solution to initially obtain the preferred topics/headings that are of importance to be included in the summary. We use existing systems and methods to generate summaries, with the novelty focusing on a domain-specific approach for commercial documents. The topics/headings from a given contract are made available to the user to choose from. This would make the generated summary accurate and caters to the unique needs of individual users. We have explored only the extractive ways to summarize a document. We have abstained from using abstractive summarization techniques as a large number of input documents are required to train a supervised model. This problem can be addressed by aggregating more input data with human-written summaries and using a supervised methodology to get better results. We look to expand on existing technologies and validate a tool for automatic summarization of legal documents that would most certainly be useful to lawyers, corporates, professionals to review

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
pp.19-26, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111202

various contracts. Even common men could potentially use it to obtain a general idea of the contracts they are about to sign or others concerning their interests. Having said that, it might not work for someone viewing a contract for the first time as they might fail to see the domain-specific importance that it carries.

2. RELATED WORK

Haghighi and Vanderwende [1] presented an exploration of generative probabilistic models for multi-document summarization. They started with a basic word frequency-based model and developed a sequence of models such as SumBasic, KL-Sum, TopicSum, and HierSum. HierSum was a hierarchical LSA-based summarizer, which gave the best ROUGE score.

Galgani et al. [2] compared traditional summarization methods with rule-based systems with a custom knowledge base and catchphrases acquired from legal documents acquired from the Federal Court of Australia. They show that the knowledge base created outperforms traditional summarization techniques.

Polsley et al. [3] proposed a tool called CaseSum for automatic text summarization of legal texts. They combined the word frequency method with additional domain-specific knowledge such as the involved parties, abbreviation of entity names. They used ROUGE as well as a custom domain expert to evaluate their approach.

Manor and Li [4] proposed a method for summarizing the Terms of Service. They tested out extractive summarization methods and compared them with human-written summaries. Their work and conclusions aligned most with our work and they are further discussed in the coming sections.

Erera et al. 2019 [5] proposed a novel method that generated summaries for research publication in the computer science domain. Each research paper was parsed from which tables, images, titles, and other metadata were extracted. Along with this they also extracted different types of entities and utilized a custom Unsupervised query focused multi-document summarization using the cross-entropy method. [6]

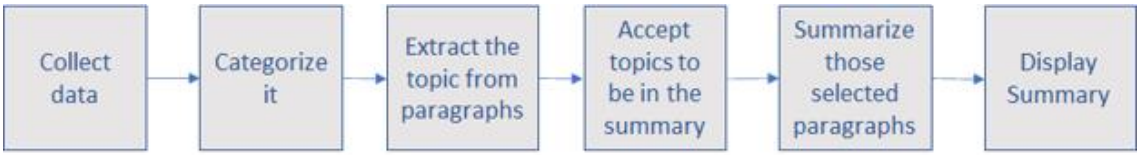


Figure 1. Workflow of the project.

3. PROPOSED METHODOLOGY

In this section, the various steps of the summarization process, as depicted in Figure 1, are discussed.

3.1. Collection of Data

The first step in building a model to summarize a text is to collect, categorize, and pre-process data. As mentioned earlier, we are considering the case of “Employment agreement”. The total number of samples collected is 1000, taken from the open-source repository of LexPredict [7].

3.2. About the data and Categorisation

As mentioned above, we are focusing on the sub-domain of the "Employment Agreement". There are two divisions for an employment agreement. One is a newly issued one, and the other is the amendments to the previous original agreement. It is observed that amendments usually contain less information. So we have our first 2 categories: "Amendments", "Agreements".

From the collected dataset, it is observed that some of the contracts are merely empty forms. So those are to be omitted. They are categorized as "Empty". As mentioned before, a majority of documents contain headings/ topics succeeded by paragraphs. Further, the "Agreements" are categorized as those with "Headings", and those with "Without Headings". Since it is important to tokenize the documents as paragraphs and further into sentences, we must know how the paragraphs are segmented. Subsequently, the documents with "Headings" are further categorized as "Alphabets", "AlphaNum", "Number.Number", "Number", and "Roman", meaning how they are indexed in the document. These are depicted in Figure 2.

The categorization is mainly done to find out how each topic/ heading is indexed so that it will be easier to extract them.

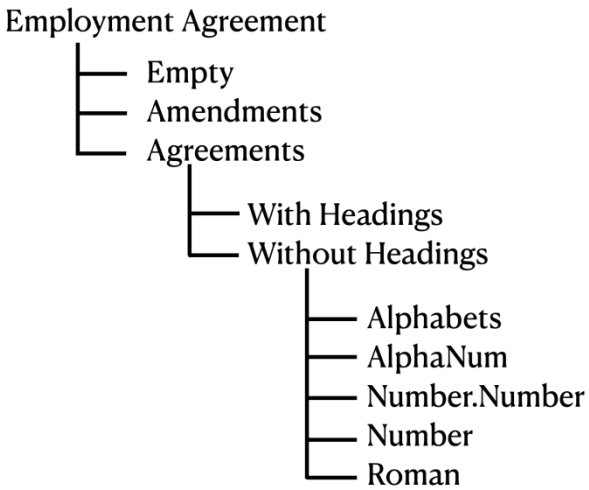


Figure 2. Categorisation of the dataset.

3.3. Data Pre-processing

Once the categorization is done, the data is subjected to cleaning and preprocessing for the task of summarization. Using basic python formatting techniques, the topic-paragraph pairs can be extracted and inserted into a dictionary. This is done for the entire document that is uploaded.

3.4. Topic Extraction

The topic extraction is based on the observation that the majority of the documents are indexed (Alphabets, AlphaNum, Number.Number, Number, Roman), and contain heading/ topic for the corresponding paragraphs, as seen in Figure 3. For the remaining documents, in the future, this project can be expanded where we can train a model to identify the topic of the paragraph and then map it with its corresponding content.

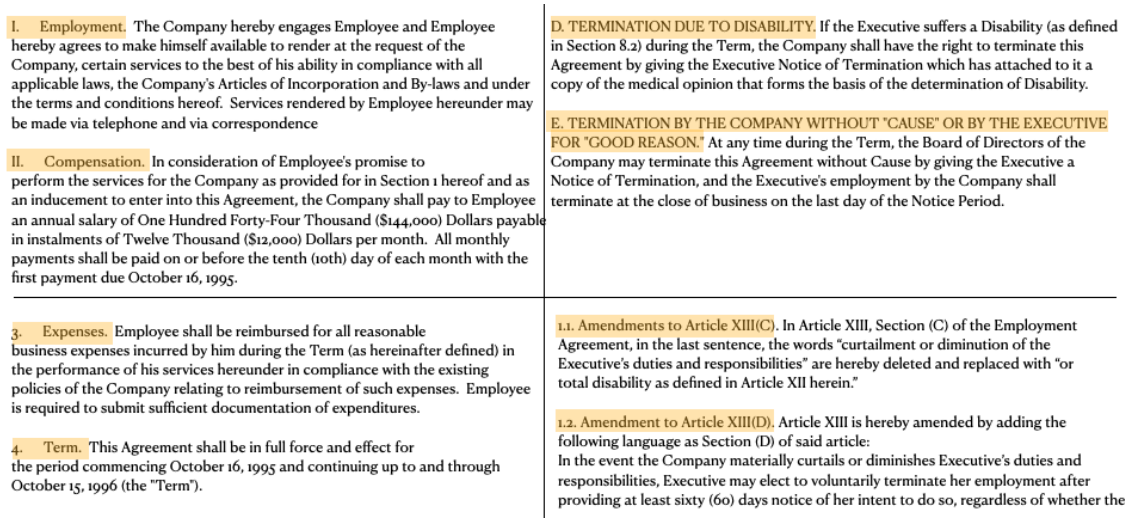


Figure 3. Some of the topics extracted from documents. From left top corner, clock-wise: Roman, Alphabets, Numbers, Number.Number

3.5. Models Used

3.5.1. Tf-Idf Summarization

Term frequency-inverse document frequency is used as a weighting factor for term features. For each term in the document, the weight increases as the word frequency increases, but it is offset by the number of times the word appears in the entire data set. The logic behind this is that if a term or word appears frequently, it's important. But if it appears frequently in other documents as well, it's probably not that important, and therefore alters its weight accordingly. This is the drawback that from using the bag-of-words model as it took into account all the frequent words without discrimination.

3.5.2. TextRank

The TextRank algorithm [8] was inspired by the famous PageRank algorithm, which models any document as a graph using sentences as nodes. It determines the relation of similarity between two sentences based on the content they both share. This overlap is calculated simply as the number of common lexical tokens between them, divided by the length of each to avoid promoting lengthy sentences.

3.5.3. LexRank

LexRank Algorithm [9] is similar to the TextRank algorithm as discussed before. It uses a modified version of the PageRank algorithm to rank the sentences in the document. It models the document as a graph using sentences as its nodes. But unlike TextRank, where all the weights are assumed as unit weights, LexRank utilizes the degrees of similarities between words and phrases. Then calculates the centrality of those sentences and assigns the weight to the node. Modified cosine similarity is then used to compare the similarity between two sentences.

3.5.4. Latent Semantic Analysis

Latent Semantic Analysis [10] is a technique that analyzes relationships between document sentences, first by constructing a document term matrix, which is a representation of each of the document sentences as vectors, where the rows correspond to the document sentences and the columns are unique words present in the vocabulary. Then Singular Value Decomposition is used to reduce the number of rows while still capturing the structure among the columns. Finally, cosine similarity is calculated between vectors formed by any two columns to determine the degree of closeness.

3.5.5. KL-Sum

Statistically speaking, KL-divergence [11] is a measurement used to find the difference between 2 distributions. KL-Sum is a greedy optimization approach that measures the divergence of the summary vocabulary words from the input document vocabulary words. It adds sentences to the summary so long as it decreases this divergence value. There are 2 main criteria for selecting a sentence to be in the final summary: The KL Divergence between the input vocabulary's set of unigrams and the output/ summary vocabulary's set of unigrams. And the number of words in the summary should be less than L. The algorithm, although is similar to PageRank and TextRank, at its core KL Sum uses the KL Divergence formula to measure how different each sentence is from one and other.

We made use of the package Sumy [12] for executing LSA, LexRank, TextRank, KL-Sum.

4. EVALUATION METRICS AND RESULTS

In this section, we discuss two ways to evaluate the generated summaries. Table 1 summarizes the evaluation results for the models used.

4.1. Rouge

Recall-Oriented Understudy for Gisting Evaluation is a set of metrics used for evaluating automatic summarization and machine translation. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. [13]

Recall in the context of ROUGE means how much of the reference summary is the system summary recovering or capturing.

$$Recall = \frac{number\ of\ overlapping\ words}{total\ words\ in\ reference\ summary}$$

Precision on the other hand measures how much of the system summary was relevant or needed.

$$Precision = \frac{number\ of\ overlapping\ words}{total\ words\ in\ system\ summary}$$

The F-measure considers both the precision and recall and is the harmonic mean of the two.

- ROUGE-N: Overlap of N-grams between the system and reference summaries.
- ROUGE-1: Refers to the overlap of unigrams (each word) between the system and reference summaries.

- ROUGE-2: Refers to the overlap of bigrams between the system and reference summaries.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics. It takes into account sentence level structure similarity naturally and identifies longest co-occuring in sequence n-grams automatically.

Table 1. F-measure scores of the 5 models used.

Model/ Metric	ROUGE-1	ROUGE-2	ROUGE-L
LexSum	0.4916	0.1898	0.4421
TextRank	0.5098	0.2366	0.5096
KLSum	0.4799	0.1745	0.3957
LSA	**0.5382**	**0.2399**	**0.5099**
Tf-Idf	0.4902	0.1908	0.4286

5. USER INTERFACE

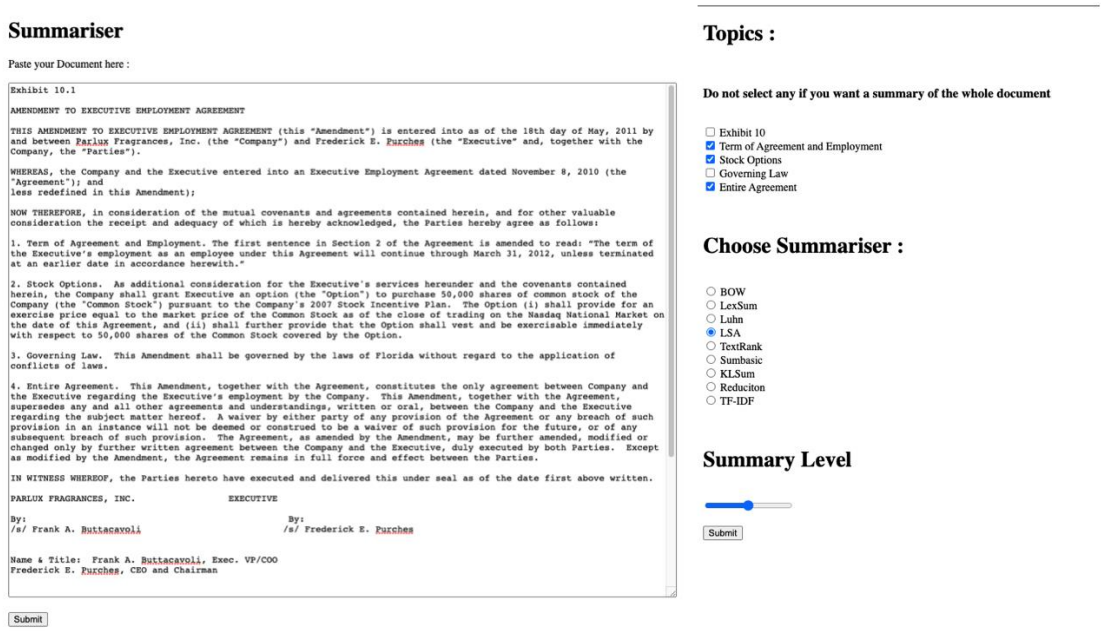
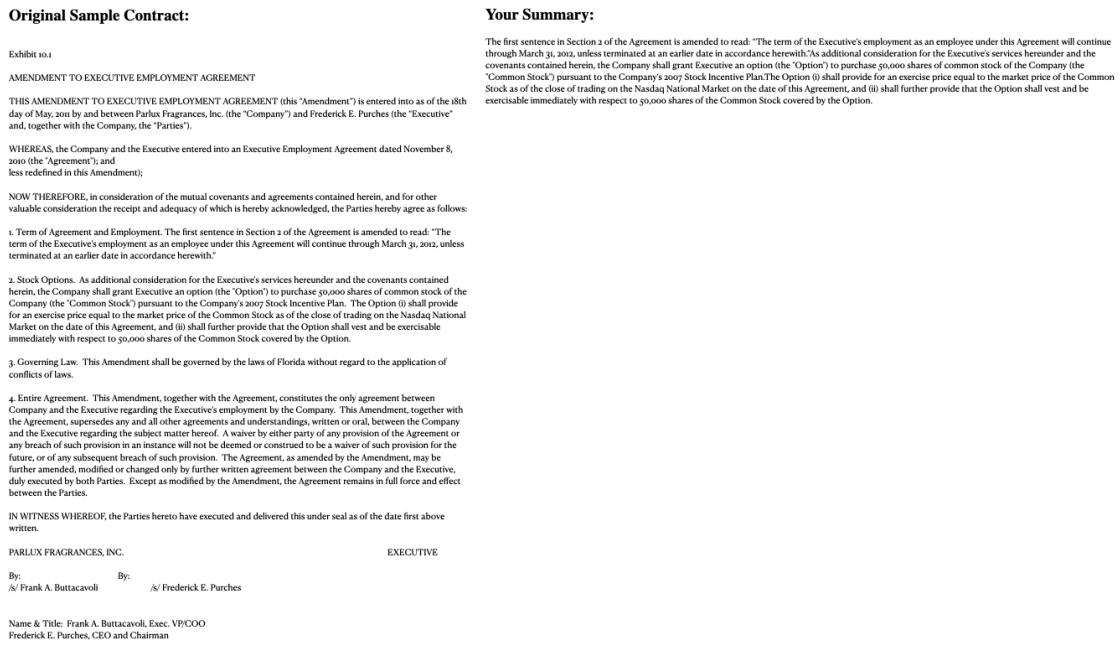


Figure 4. Uploading contracts and Topic extraction.

On the left-hand side of Figure 4, a sample employment agreement is uploaded. On the right-hand side of Figure 4, the topics are extracted and displayed to the user. The user selects the topics that are to be included in the summary and how detailed the summary has to be. (Choosing the Summarizer is for the paper's explanation point of view).

The summary of the uploaded contract is displayed in Figure 5. For the sake of simplicity, the best performing LSA is chosen to summarize the input document.



Original Sample Contract:

Exhibit 10.1

AMENDMENT TO EXECUTIVE EMPLOYMENT AGREEMENT

THIS AMENDMENT TO EXECUTIVE EMPLOYMENT AGREEMENT (this "Amendment") is entered into as of the 18th day of May, 2011 by and between Parlux Fragrances, Inc. (the "Company") and Frederick E. Purches (the "Executive" and, together with the Company, the "Parties").

WHEREAS, the Company and the Executive entered into an Executive Employment Agreement dated November 8, 2010 (the "Agreement"); and
less redefined in this Amendment);

NOW THEREFORE, in consideration of the mutual covenants and agreements contained herein, and for other valuable consideration the receipt and adequacy of which is hereby acknowledged, the Parties hereby agree as follows:

1. Term of Agreement and Employment. The first sentence in Section 2 of the Agreement is amended to read: "The term of the Executive's employment as an employee under this Agreement will continue through March 31, 2012, unless terminated at an earlier date in accordance herewith."

2. Stock Options. As additional consideration for the Executive's services hereunder and the covenants contained herein, the Company shall grant Executive an option (the "Option") to purchase 50,000 shares of common stock of the Company (the "Common Stock") pursuant to the Company's 2007 Stock Incentive Plan. The Option (i) shall provide for an exercise price equal to the market price of the Common Stock as of the close of trading on the Nasdaq National Market on the date of this Agreement, and (ii) shall further provide that the Option shall vest and be exercisable immediately with respect to 30,000 shares of the Common Stock covered by the Option.

3. Governing Law. This Amendment shall be governed by the laws of Florida without regard to the application of conflicts of laws.

4. Entire Agreement. This Amendment, together with the Agreement, constitutes the only agreement between Company and the Executive regarding the Executive's employment by the Company. This Amendment, together with the Agreement, supersedes any and all other agreements and understandings, written or oral, between the Company and the Executive regarding the subject matter hereof. A waiver by either party of any provision of the Agreement or any breach of such provision in an instance will not be deemed or construed to be a waiver of such provision for the future, or of any subsequent breach of such provision. The Agreement, as amended by the Amendment, may be further amended, modified or changed only by further written agreement between the Company and the Executive, duly executed by both Parties. Except as modified by the Amendment, the Agreement remains in full force and effect between the Parties.

IN WITNESS WHEREOF, the Parties hereto have executed and delivered this under seal as of the date first above written.

PARLUX FRAGRANCES, INC. EXECUTIVE

By: By:
/s/ Frank A. Buttacavoli /s/ Frederick E. Purches

Name & Title: Frank A. Buttacavoli, Exec. VP/COO
Frederick E. Purches, CEO and Chairman

Your Summary:

The first sentence in Section 2 of the Agreement is amended to read: "The term of the Executive's employment as an employee under this Agreement will continue through March 31, 2012, unless terminated at an earlier date in accordance herewith."As additional consideration for the Executive's services hereunder and the covenants contained herein, the Company shall grant Executive an option (the "Option") to purchase 50,000 shares of common stock of the Company (the "Common Stock") pursuant to the Company's 2007 Stock Incentive Plan.The Option (i) shall provide for an exercise price equal to the market price of the Common Stock as of the close of trading on the Nasdaq National Market on the date of this Agreement, and (ii) shall further provide that the Option shall vest and be exercisable immediately with respect to 30,000 shares of the Common Stock covered by the Option.

Figure 5. The original document (left), and the corresponding generated summary (right).

6. DISCUSSION AND CONCLUSION

From our results, we conclude that the summarization of legal/ commercial documents is a challenging task and could further be improved. From Table 1, we see that the F-measure scores for each of the extractive summarization models are satisfactory. LSA performs the best amongst others. LSA captures both the meaning of words as well as the similarity among the sentences. Also, Singular Value Decomposition (SVD) can reduce noise and model latent, the semantic relationship among words and sentences. This leads to an improvement in accuracy. The reason this is a challenging task is that firstly, the formatting and the representation adapted companies to draft legal documents to vary hugely, hence the task to text pre-processing is difficult. Second, the use of current SOTA supervised or unsupervised models for text summarization will fail to work because it is difficult for it to recognize legal jargon and taxonomy. As [4] rightly mentions there is no large dataset available for this domain. This task could further be attempted to solve by training a supervised abstractive summarization model, using Neural networks. This, of course, requires a large number of documents and their corresponding human-written summaries.

REFERENCES

[1] Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-Document Summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Boulder, Colorado, 362–370. https://www.aclweb.org/anthology/N09-1041

[2] Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. 115–123.

[3] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. The COLING 2016 Organizing Committee, Osaka, Japan, 258–262. https://www.aclweb.org/anthology/C16-2054

[4] Laura Manor and Junyi Jessy Li. 2019. Plain English Summarization of Contracts. In Proceedings of the Natural Legal Language Processing Workshop 2019. Association for Computational Linguistics, Minneapolis, Minnesota, 1–11. https://doi.org/10.18653/v1/W19-2201

[5] haiErera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Nakash, OdelliaBoni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, AchiyaJerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and David Konopnicki. 2019. A Summarization System for Scientific Documents.

[6] Guy Feigenblat, Haggai Roitman, OdelliaBoni, and David Konopnicki. 2017. Unsupervised queryfocused multi-document summarization using the cross entropy method. In Proceedings of the 40th International ACM SIGIR, pages 961–964.

[7] LexPredict, LLC, acquired by Elevate Services, Inc. in 2018. https://github.com/LexPredict/lexpredict-contraxsuite-samples

[8] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://www.aclweb.org/anthology/W04-3252

[9] Gunes Erkan and Dragomir Radev. 2011. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. Journal of Artificial Intelligence Research - JAIR 22 (09 2011). https://doi.org/10.1613/jair.1523

[10] Makbule Ozsoy, Ferda Alpaslan, and Ilyas Cicekli. 2011. Text summarization using Latent Semantic Analysis. J. Information Science 37 (08 2011), 405–417. https://doi.org/10.1177/0165551511408848

[11] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. The Annals of Mathematical Statistics (1951), 79–86.

[12] https://pypi.org/project/sumy/

[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Tex Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

RESEARCHING BLOCKCHAIN TECHNOLOGY AND ITS USEFULNESS IN HIGHER EDUCATION

Shankar Subramanian Iyer[1], Arumugam Seetharaman[2] and Bhanu Ranjan[3]

[1]Researcher, S.P. Jain School of Global Management, Dubai
[2]Dean Academic Affairs at S P Jain School of Global Management, Singapore
[3] Associate Professor, S.P. Jain School of Global Management- Singapore

ABSTRACT

The current paper focuses on the potential of using Blockchain Technology (BCT) in the Higher Education Domain and explores its usefulness in solving Higher Education issues. This research discusses the Blockchain features, challenges and its benefits in education, followed by review of some current Blockchain Higher Education applications. This paper reviews the Blockchain Technology (BCT) and its implementation in Higher Education. This research used a quantitative methodology and stratified clustered simple random sampling approach. Data has been gathered through an online survey instrument and the partial least squares structural equation modelling (PLS-SEM) technique applied to 383 responses. Blockchain technology has its unique features, benefits that can solve Education system requirements, and its successful implementation issues discussed. An effort made to gather enough consensus to build future implementation. The integrated model of Blockchain features matched to the needs of the Education System by agreement of the experts (discussions), and a survey conducted involving the students, teachers, educationists, Blockchain experts, and professionals, is tested and validated by SEM using PLS.

KEYWORDS

Blockchain Technology (BCT), Higher Education Implementation, Higher Education Domain, Higher Education Management, Higher Education Technology, Structural Equation Model.

1. INTRODUCTION AND BACKGROUND

Blockchain Technology is known to most people about Bitcoin and other cryptocurrencies (Zhao et al., 2016). Blockchain has been now around for about a decade, other than in cryptocurrency it has not been very successful. It has been applied to many areas of Business-like finance, judiciary, Higher education, healthcare, logistics, and commerce; however, with limited success. In its current form in these industries, it can be termed more to be a smart system than a Blockchain. Research has shown that many small implementations in Higher Education have happened; however, success is limited. The full potential of Blockchain Technology in education is desired. The Blockchain Technology features and benefits proven so far have the potential to address most of the challenges currently experienced in Higher Education Framework (Alam et al.,2020), (Mahmood et al., 2020).

The rigidity of the current Higher Education System prohibits the Learner from choosing what to study, in terms of focus on a specific topic. Blockchain Technology is suitable for the Higher Education Domain due to its immutability, transparency, and trustworthiness characteristics, which can be useful in Higher Education application (Underwood, 2016).

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.27-48, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111203

Exploring Blockchain Technology: The Security feature of Blockchain is valuable to the Higher Education sector due to its digital signature and encryption. The Blockchain Technology system for Higher Education needs to be secure, convenient, and tamperproof to keep records of certification and transcripts. Blockchain Technology can provide a system which can control frauds (Chen et al., 2018). The system based on data stored in several decentralized ledgers and encrypted makes it difficult for hackers to get access. Information, if lost, can be quickly recovered. This feature is very crucial in Higher Education as the records need to be secured. Blockchain Technology maintains transparency. The Participants of transactions notified about the completion of transactions, which is both convenient and trustworthy (Alexopoulos et al., 2019). It is free of Intermediates and so no hidden fees as the system is decentralized, free of fees and faster settlements. The Access levels have to be decided by the Users where access is available to anyone (Public) or authorized permission given for each node (restricted) (Swan, 2015). The Speed of Transactions is processed much faster than usual, as there is no need to include payment systems, which reduce the cost and increases the processing speed. Since Account reconciliation is immediate, the validity of transactions is checked and confirmed by participants, thus leading to authenticity. All of these features of Blockchain meets the requirement in the Higher Education Framework (Zheng et al., 2017). Blockchain Technology experience tells us that if there has to be a Blockchain revolution, many barriers- technological, governance, organizational and even societal- will have to break. It would not be proper to proceed with Blockchain implementation without understanding how it is likely to influence the Higher Education field (Lakhani, 2017); (Hughes et al., 2019). It is of interest to understand private Blockchain networks which have been around for some time; however, not integrated (Zheng et al., 2018).

Blockchain Technology usefulness to Higher Education: Most Learners have issues to get their old Higher Educational certificate authenticated, acquired a long time back. The hard copy submitted to the Employer for employment or for pursuing higher studies usually needs authentication by the relevant authorities like the HR Department, the College Administrators, the Ministry of External Affairs. The current system has the issue of a long wait; it would take more than 3- 4 weeks provided the Certificate issuer still exists. If the college from which the Learner has graduated does not exist now, the same certificates look suspicious to the current people who go through it. Such incidents cannot occur if the records are maintained on a Blockchain system, as the record is maintained in different Ledgers all separately kept. In this article, an overview on the BCT with required details is discussed that apply to integrate individual institutions at local level, groups of institutions at the national level to common Blockchain Higher Education platform (Lizcano et al., 2019). Some of the applications are school records management by maintaining verifiable Student transcript and degrees which can be transferred to remote storage easily (Chen et al., 2018). The student privacy, confidentiality is ensured by authorized permissions and management of records using BCT. The public funds distribution and private grants given, student loans payments; license/dissertation/PhD thesis management is easily tracked using the BCT. If the Blockchain Technology integrated across the various sectors and Industries, it would revolutionize the use of BCT. The objectives of this paper are to give insights on the use of Blockchain Technology for Higher Education applications, to highlight the state-of-the-art techniques that currently used to provide these services, to examine their challenges, and to discuss how the Blockchain Technology can resolve these challenges (Batubara et al., 2018).

Blockchain Technology usefulness to Higher Education is undeniable (Holotescu, 2018).

2. RESEARCH PROBLEM

Research Questions

a. Are the features of Blockchain Technology suitable to Higher Education?
b. Can the Blockchain Technology features & benefits be useful in Higher Education Framework?

3. OBJECTIVES OF THE PROPOSED RESEARCH

a. Investigate the suitability of the features, the usefulness of Blockchain to Higher Education Sector

b. Review the benefits & features of using Blockchain Technology in Higher Education Framework and Investigate the future of Blockchain in Higher Education.

4. SCOPE OF THE STUDY

This Research proposes to conduct a literature review of the usefulness of Blockchain Technology to various sectors and how it can be used in Higher Education to ensure authenticity, avoid frauds, ensure data security. Research until today has shown that the implementation of Blockchain Technology in most sectors, including Higher Education, has been limited and not utilized the immense potential of the BCT. The Research article is to find the gaps and to see the ways all the features, benefits utilized to make it commercially viable to Higher Education Framework. The working of the Blockchain Technology giving the technical details is not in the scope of this paper.

5. PAPER ORGANIZATION AND THE RESEARCH BASIS

This paper organizes the references of various researches done earlier and derive the Literature Review, explaining the purpose for the same. The Gap Analysis developed by identifying the gaps mentioned in the Research papers studied. From this, the most frequently occurring gaps taken for study purpose for this paper. The dependent variable identified as the "Usefulness of Blockchain Technology in Higher Education" and the factors which frequently occur in the above-identified articles for papers are listed. The six factors with the highest frequency form the independent factors to determine the conceptual model shown in figure 1. Based on the above Research problems, research objectives and the possible solutions are determined to form the Research Framework.

Then the Usefulness of the Blockchain Technology to the Higher Education is discussed. The benefits and features of Blockchain Technology have been listed and explained in details. Next, the challenges in the implementation of the same and the hurdles anticipated discussed. The possible solutions suggested in the implementation and the Methodology that followed discussed: the limitations and the conclusion given at the end. The references have been cited at each juncture to make it more creditable and make the arguments robust.

6. SURVEY OF LITERATURE

6.1. Purpose of the Literature Review:

The purpose of the Literature review is to collect information and knowledge on the Blockchain and its various applications and focus on Higher Education (Rowley et al., 2004). The areas of prior studies identified to prevent duplication of the work and to give credit to the other Researchers wherever their material used. It helps to look for inconsistencies, gaps in Research, conflicts or open questions left out in the earlier researchers. The need for additional Research identified and justified. The above would be the contribution to the topic and justify the further study needed. The other research papers and literature study will help the Research with ideas, conclusions and theories, establish similarities and differences and notice principal methodologies and research techniques used (Risius et al., 2017). Most Researches have used secondary Research to come to their conclusions. First, the Articles identified and screened before being included in the Literature Review. The keywords used by the Authors help in identifying the relevant articles. The Keywords mentioned above used in Google Scholar, Mendeley, Article Publishers website like Scopus, Harvard Review, IEEE for the above purpose (Firdaus et al., 2019). It has to be decided what articles to consider for review and which ones to exclude. The main reason is to exclude articles or topics, which are out of the scope of the study, like Bitcoin and Cryptocurrencies and discussing trading (Hart, 2018). It helps to understand the style of writing and research methodology followed by significant researchers in Blockchain and its application in Higher Education, which is majorly secondary research, and Inductive. Therefore, it led the way to make the Research Plan (Ahmed et al., 2018).

6.2. Detailed explanation as to how the literature was identified

The Secondary Research using the Keywords and systematic search and review, 110 articles were identified from credible journals like Scopus, IEEE Access, ProQuest, library resources, upward of 2016. Only SCJ used wherein if the H Index is more than 60, then the journal is considered credible. Care to restrict the literature review search only to peer-reviewed articles so that the selected articles or literature is credible. Sufficient attention to check currency, relevance and reference of the Articles. Seventy-seven articles identified using the inclusion approach and balance excluded due to a lack of quality material to using the Keywords and systematic search and review, 110 articles were identified from credible journals like Scopus, IEEE Access, ProQuest, library resources, upward of 2016. Only SCJ used wherein if the H Index is more than 60, then the journal is considered credible. Sufficient care to restrict the literature review search only to peer-reviewed articles so that the selected articles or literature is credible. Sufficient attention taken to check, currency, relevance and reference of the Articles—seventy-seven articles identified using the inclusion approach and balance excluded due to a lack of quality material used. The articles then compiled into a Summary table identifying the significant findings on which the variable depend on (Machi et al., 2016). The table includes the type of Research done, the Primary and Secondary Research, which has used to collect data and to analyze. The limitations or gaps made to produce the Gap Analysis and the Research conceptual model shown in figure 1. The whole idea is to identify credible references for putting forward the topic attributes in a systematic manner (Creswell et al., 2017). The above will avoid duplication, and it gives the readers a synopsis of the things to come and motivates them. The type of data used by the major of them is secondary data through conference papers, journal reviews, government data, grey literature, books. So, the idea to go further with our Research design ascertained.

The significant findings with the dependent, independent, sub-variables identified, and the frequency of occurrence will help to ascertain Research Variables (Fink, 2019). It also enables the Researcher to focus on contribution to the existing Research Topic like Energy Requirements in the Blockchain Technology and the Regulator, First Sponsor necessity in today's context is the contribution of this Research Article (Koteska et al., 2017). This article should look at the perspectives to be included and excluded and should be reflective (Miles et al., 1994), (Batubara et al., 2018).

6.3. Usefulness of Blockchain in Higher Education

From the Literature review, it has identified that the Independent Constructs are Decentralization, Traceability, Immutability, Currency properties, Scalability, First Sponsor Organization and Energy Requirements. These are the main features of Blockchain Technology, which will be useful in Higher Education Domain, identified as the Independent Variable (Johnson, 2018); (Beck, 2018). Most of the Systems are centralized, i.e. all the data are maintained in a single spot, and this makes it vulnerable to the risk of losing data, being hacked or the information compromised, for example, Facebook data leakage and Education data leakage from Universities, Education websites (Cheng et al., 2017), (Alneyadi et al., 2016).

Decentralization is the process of keeping Data in various places and not with a single entity. In Blockchain, in peer network with decentralized structure the system has various nodes with storage, data verification, maintenance, and transmission facilities. A decentralized system, built trust using mathematical methods between distributed nodes. In Higher Education, it translates to storing certificates, records of Learners, maintained and transmitted on Blockchain system with lesser fear of being hacked or compromised (Huang et al., 2017). The Blockchain uses cryptic codes to encrypt transactions, done by Miners, which is complex to crack and gets more complicated with increased transactions and nodes added to the network. It does not involve any third-party intermediator and Employers can verify Student certificates at the click of the mouse. (Chen, 2018)

Traceability means to be able to trace the source from the existing position. For example, if potential Employee Profile is claiming to be Masters, Blockchain can trace the source of the Master's certificate, issued by which University, the transcript and the Mark sheet. Blockchain Technology promises traceability, provenance and transparency of information. At the same time, there is reassurance that Participants cannot change the information in a bid to hide the exact origin without consensus from each of them (Khachaturova et al., 2018). Traceability is possible by linking block information through hash keys. All transactions need to be arranged chronologically on Blockchain, and adjacent blocks are connected using the hash function cryptographically (Salman et al., 2018).

Immutability is an integral part of the Blockchain due to its structure, as all transactions storing blocks are linked using hash keys from the previous block and another hash key linking to the next. In the Higher Education system, every new transaction needs to be linked to an old block and made very secure, thereby avoiding system manipulation. Tampering in an evolved Blockchain is possible only if 51% of the ledgers stored by the network is changed which is very unlikely to happen (Tschorsch et al., 2016), however, can be an inherence for small networks.

Currency Properties: Blockchain Technology in Higher Education and tokens/rewards like Taelim coins go together; every Blockchain network has a potential cryptocurrency property. Point- to- point transactions are the essence of Blockchain Technology with no third-party involvement. The Tokens or coins meant to compensate the Miners for encrypting each transaction to be recorded. The above means lot of energy requirements, more than used in smart

systems. Therefore, the Miners have to reward for their efforts and energy usage (Panda et al., 2019).

Scalability: Most Blockchains have problems when the transactions increase beyond some particular level, and the whole system slows down. There is need to work on Increased Capacity, better known as scalability to handle and make transactions faster using the Miners/Nodes. The miners get rewarded with token coins. Delegated proof-of-stake is a viable consensus mechanism by which users, vote on a small number of delegates, who maintain the ledger, in this case, would be identified Miners. The reduction in active nodes means the network can increase its throughput. Each node is paid through inflation and can justify running a large data center to support the network. It depends on the Segwit, Block Size Increase, Sharding, Proof of Stake, Off-Chain State Channels and Plasma used (Salman et al., 2018).

First Sponsor Organization: Most people view not having a Regulator as the main advantage of Blockchain Technology. However, it is the major reason financial systems do not accept it. First Sponsor Organization like UAE Smart Government can be the guarantor of sorts; they take care of the regulations required to set it up and give its credibility in financial parlance. It ensures the credibility of the members, the miners, the ledger keepers who identified to avoid money laundering, illicit money flows (Bryson et al., 2017).

Energy Requirements: The Blockchain Technology needs much energy to create the codes, to record, store and to transmit these across networks (Nakamoto, 2008). The energy consumption is much higher than used by the smart system of similar magnitude. Mining consumes a lot of energy, which has been a limitation in it spreading fast across the sectors and to countries (Truby, 2018) ;(Tschorsch et al., 2016); (Fanning, 2016).

Technology Adaption Model and Acceptance model explains the way users accept new technologies like AI, ML, Blockchain over some time. The perception changes with the use of technology to start accepting and take advantage of Blockchain. The initial resistance and mindset need to be managed with the spread of awareness and induce the usage of modern technologies (Verma et al., 2016); (Wu et al., 2017).

7. RESEARCH FRAMEWORK –

Suggested Solution Framework Blockchain in Higher Education Domain, Usefulness of Blockchain in Higher Education as shown in figure 1.

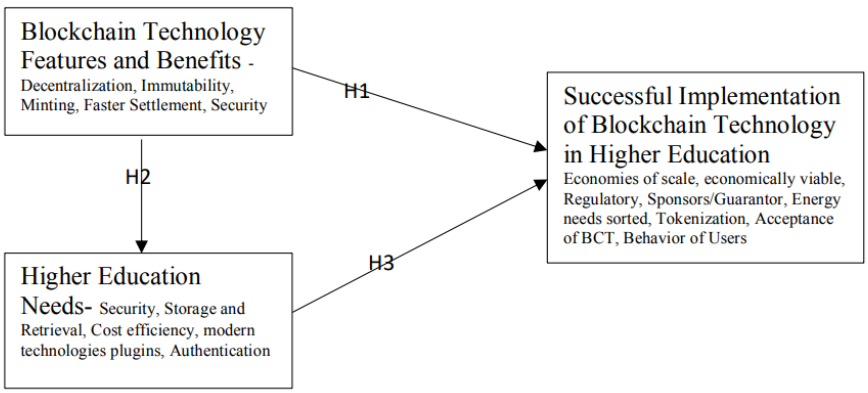
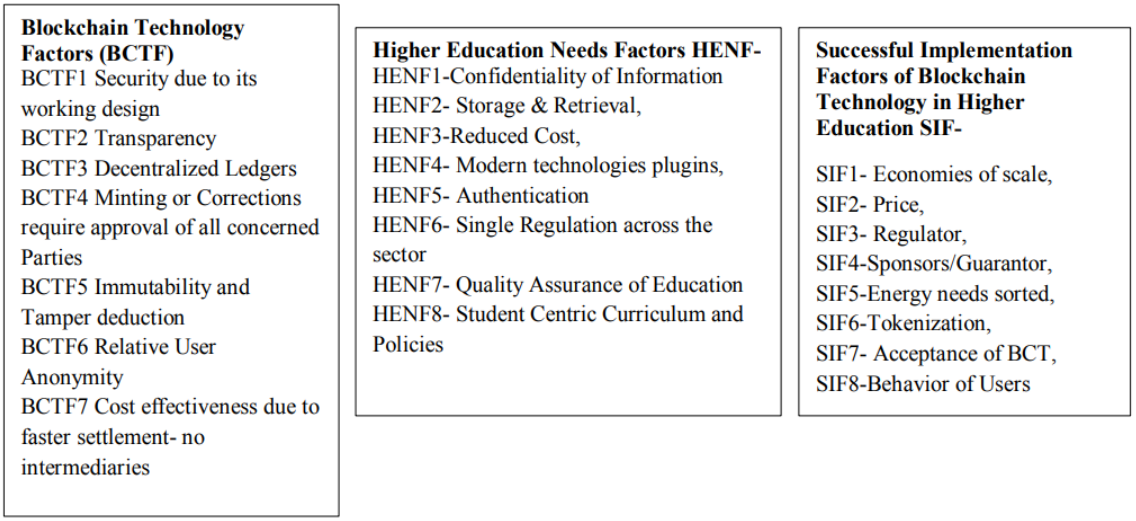


Figure 1

H1: The Reliability, Immutability features of Blockchain be useful for successful implementation of Blockchain Technology in Education?

H2: The trust and efficiency features of Blockchain Technology be useful for storing records of Education.

H3: The Higher Education needs to be addressed for successful implementation of Blockchain in Higher Education



8. RESEARCH METHODOLOGY

8.1. Research Need and type of Research required

Blockchain has been most discussed technology in the last four years. Many things discussed Blockchain and its features are helpful; however, the way to implement Blockchain has been missed. Therefore, it is a matter of interest to research the lack of implementation, the issues, and the challenges to its implementation, especially to Higher Education. The technology development in Blockchain initially experimented using cryptocurrency, and the experience has been maximum in this field. Therefore, it is the root of all Blockchain development and slowly limited successes in other sectors, Banking, Finance, Governance, Higher Education, Health care, Logistics (Johnson et al., 2019).

The Research Topic we endeavor to study of the current status of Blockchain and its implementation in various sectors, study the usefulness of the features of Blockchain to Higher Education Domain, the usage in various sectors, the challenges, the sectors in making it widely applicable in Higher Education Domain. We will be using both the types of Research based on the nature of information needed from experts and through survey, interviews (Mertler, 2018).

8.2. Primary Research conducted through Online surveys

Quantitative Data (300 plus Sample Size planned). The researcher sends a communication to 835 target respondents across the various cities of India, Malaysia, Singapore, UAE. The survey is achievable in a Smart Government environment like in Dubai, for example, where 90,000 government Employees from 28 entities are on the same platform. For the awareness, survey Questionnaire on Blockchain and its application to Higher Education was sent to 835 respondents

across various Government agencies like Energy companies, Transportation authorities, Education development agencies, Higher Education Ministry, Universities and Colleges in UAE, India, Singapore, Malaysia, Pakistan, affiliated to the Higher Education Ministry (Mertler, 2018).

8.3. Secondary Research

Deterministic Research and accepted factors included in the Research and contribute to the gap identified from various Research papers studied. Correlation analysis used to see how the dependent variable is actually related to the independent Variable. Try to see how many of the independent variables identified have a strong influence on the adaption of the Blockchain Technology. For example, how much influence a Regulator or the First Sponsor/Investor Organization has on the adaption of Blockchain Technology in the Higher Education Domain. Expert opinion from the People in the industry in Higher Education assimilated, or from Consultants. Users in the Blockchain Industry from Banking, Finance, Regulators, Students, Researchers identified for this purpose- Qualitative Data (12 Experts) (Quinlan et al., 2019).

The various philosophies, the Positivism approach seems to be useful for Research for considering the Quantitative data through surveys followed by the analysis using correlation, regression using SPSS software and Adanco. It is required to interview, conduct focus groups for getting the opinion of the experts about the Blockchain Technology application to the Higher Education Framework. It will help to create the Gap Analysis chart, Research Framework by identifying the dependent Variable, Independent Variable and the sub-variables to establish the relationship between them (Machi et al., 2016).

9. DATA COLLECTION

9.1. Research Approach

A quantitative research methodology uses a deductive research, uses structured approach, statistics, and a large adequate sample size to analyze data to come to conclusions. The Sampling population of this Research is Blockchain professionals, Educationists, Blockchain Users and Students working remotely due to forced lockdowns due to COVID-19. Hence this study uses stratified clustered random convenient sampling methodology, which focusses on affirms that every potential working professional and student working-remotely has an equal opportunity to participate in this research (Zikmund et al., 2013). Raosoft Sample size calculator can be used to know the tentative sample size based on the confidence level 95%, margin of error-5%, is found to be 385 respondents (Raosoft Inc, 2020).

The usage and awareness of Blockchain technology are low, the ‘scenario’ method was used to convey the use-case of Blockchain application. The questionnaire for the online survey was adapted from existing questionnaires available for validating task technology fit model and technology acceptance model by using statistical methods. The list of questions for individual constructs of the research framework is shared in Annexure-I. A total of seventeen questions were asked to respondents, which measured their attitude via a 5-point Likert scale (Nemoto&Beglar, 2014). The questionnaire was formulated with consensus with discussion and pilot survey shared with 20 respondents and valuable feedback in improving the questionnaire to avoid ambiguity and bias.

SPSS used in this research, to measure demographics characteristics of the respondents and its relation with the Blockchain in Education. The widely used PLS-SEM method is used for this business research for hypothesis testing and analyzing reflective measurement and structured measurements. Reflective measurements cover indicator reliability, construct reliability,

convergent validity, and discriminant validity. Structural measurements cover predictive relevance, the significance of path coefficients and overall variance of a structured model-Results shown in Table 3.

9.2. Data validation and analysis

Respondent's characteristics

Table 1 show the demographics of participants of the online survey. A total of 835 survey questionnaires were distributed to students, and working professionals in India, Malaysia, UAE, Singapore and several other countries clubbed under the 'rest of the world.' A total of 383 respondents participated in this research survey. (see table 1)

Table 1: Demographics of Respondents

Demographic Variable	Category	Percentage
Age Group	18-25	18.24
	26-35	32.56
	36-45	27.79
	46-60	18.92
	60+	2.49
Gender	Male	55.43
	Female	44.57

Demographic Variable	Category	Percentage	Demographic Variable	Category	Percentage
Education	Highschool	1.28	Region	India	33.97
	Undergraduate	21.54		UAE	37.60
	Post Graduate	60.34		Malaysia & Singapore	15.82
	Doctoral	16.84		Rest of the World	12.61

Demographic Variable	Category	Percentage	Demographic Variable	Category	Percentage
Awareness of the Blockchain Application in Education?	Extremely familiar-Expert in the field	19.60	Association with Blockchain Technology?	Researcher	10.80
				Student/Learner	28.30
				Working IT Professional	10.24
				Business Owner	8.37
				Project Manager	7.11
				Consultant	5.76
				Government Official	8.60
	Very Familiar-working on the Blockchain Application	36.80		Regulator	2.10
				Public	6.80
	Somewhat Familiar-only researching and yet to work on Blockchain	43.60		Trader	3.50
				Miner	2.70
				Others	5.72

Table 2: Indicator Loadings

Indicator	Blockchain Features	Higher Education Needs	Successful Implementation Factors
BCTF1	0.9326		
BCTF2	0.9190		
BCTF3	0.9223		
BCTF4	0.8946		
BCTF5	0.8997		
BCTF6	0.8631		
BCTF7	0.8697		
BCTF8	0.7662		
HENF1		0.9094	
HENF2		0.9546	
HENF3		0.8676	
HENF4		0.9176	
HENF5		0.9032	
HENF6		0.8469	
HENF7		0.9349	
HENF8		0.8084	
SIF1			0.7978
SIF2			0.8142
SIF3			0.8638
SIF4			0.8817
SIF5			0.7864
SIF6			0.8149
SIF7			0.7507
SIF8			0.8852

Table 3: Construct reliability

Construct	Dijkstra-Henseler's rho (ρ_A)	Jöreskog's rho (ρ_c)	Cronbach's alpha(α)
Blockchain Features	0.9604	0.9664	0.9598
Higher Education Needs	0.9649	0.9695	0.9637
Successful Implementation factors	0.9369	0.9447	0.9328

Table 4: Convergent validity

Construct	Average variance extracted (AVE)
Blockchain Features	0.7829
Higher Education Needs	0.7992
SuccessfulImplementation Requirements	0.6816

Table 5: Discriminant validity

Construct	Blockchain Features	Higher Education Needs	Successful Implementation Requirement
Blockchain Features	**0.6810**		
Higher Education Needs	0.6260	**0.7456**	
Successful Implementation Requirement	0.6133	0.6495	**0.6928**

Table 6: Indicator collinearity

Indicator	Blockchain Features	Higher Education Needs	Successful Implementation Factors
BCTF1	3.2989		
BCTF2	4.0849		
BCTF3	3.2729		
BCTF4	4.9657		
BCTF5	3.4490		
BCTF6	4.0599		
BCTF7	4.0323		
BCTF8	2.7513		
HENF1		4.7696	
HENF2		4.8075	
HENF3		3.6898	
HENF4		4.8264	
HENF5		4.7303	
HENF6		3.9840	
HENF7		4.4588	
HENF8		3.2402	
SIF1			2.6375
SIF2			4.1317
SIF3			3.8795
SIF4			4.1534
SIF5			3.1509
SIF6			4.6817
SIF7			3.7082
SIF8			3.7229
Variance inflation factors (VIF)			

Table 7: Coefficient determination

Construct	Coefficient of determination (R^2)	Adjusted R^2
Higher Education Needs	0.7950	0.7944
Successful Implementation Factors	0.5548	0.5525

Table 8: Bootstrap direct effects inference

Effect	Standard bootstrap results p-value (2- sided)
Blockchain Features -> Higher Education Needs	0.0000
Blockchain Features -> SIF1	0.0000
Higher Education Needs -> SIF1	0.0052

Table 9: Path coefficient

Independent variable	Dependent variable	
	Higher Education Needs	Successful Implementation Factors
Blockchain Features	0.8916	0.4498
Higher Education Needs		0.3154

Based on the results from Table 9, the path coefficient is displayed below

Table 10: Path coefficient

Independent variable	Dependent variable	
	Higher Education Needs	Successful Implementation Factors
Blockchain Features	0.8916	0.4498
Higher Education Needs		0.3154

Table 11: Direct effect Inferences

Effect	Original coefficient	Standard bootstrap results					Percentile bootstrap quantiles			
		Mean value	Standard error	t-value	p-value (2-sided)	p-value (1-sided)	0.5%	2.5%	97.5%	99.5%
Blockchain Features -> Higher Education Needs	0.8916	0.8913	0.0126	70.4966	0.0000	0.0000	0.8541	0.8646	0.9143	0.9205
Blockchain Features -> SIF1	0.4498	0.4479	0.1064	4.2278	0.0000	0.0000	0.1787	0.2398	0.6596	0.7211
Higher Education Needs -> SIF1	0.3154	0.3173	0.1128	2.7952	0.0052	0.0026	0.0265	0.0949	0.5372	0.5932

10. DISCUSSION, FINDINGS AND ANALYSIS

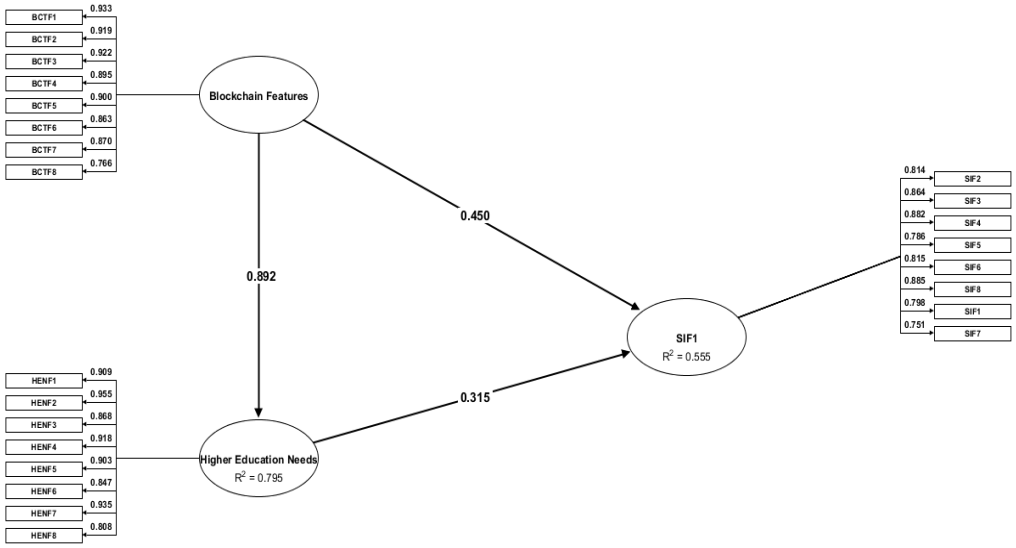


Figure 2

10.1. Discussion and Analysis

Based on the Research Final Model (Figure 2), the Model is tested and validated as per the parameters shown in the tables 1 to 10. The Hypotheses 1 to 3 have been proven to be significantly valid and acceptable. So, we accept the hypotheses and the results will prove that the Blockchain features can adequately meet the requirements of the Education system and the successful implementation is very likely.

10.2. Status of findings

H1, H2, H3 hypotheses have been accepted as the model path coefficients and the R^2 is above the acceptable value, which calculated to be 0.5555 and the hypotheses are supported and have positive significance as shown in the figure 3.

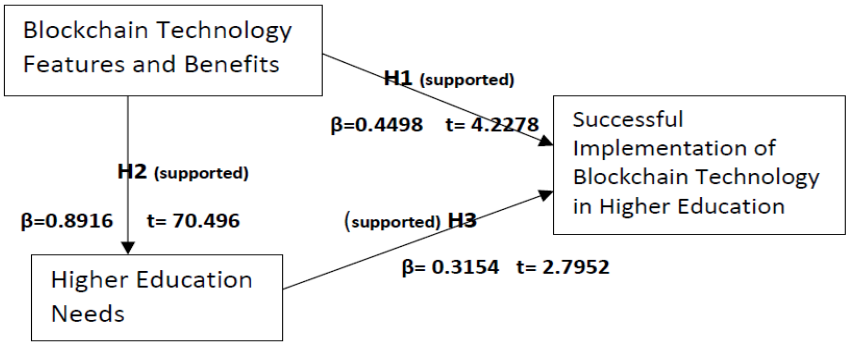


Figure 3 – the PLS-SEM validation and reliability test output

Figure 3

10.3. Features of Blockchain Technology

Decentralization, Traceability, Immutability, Currency properties, Scalability, First Sponsor Organization, Energy Requirements are the main features of Blockchain Technology, which will be useful in Higher Education Domain (Beck, 2018). The Features of the Blockchain have been discussed in the above paragraph.

10.4. Advantages of Blockchain technology, in Higher Education Framework

Some of the advantages of using Blockchain Technology in education can be (Han, 2018)

1. Reliability: Blockchain uses the decentralized ledger distributed over the various nodes in the network so the databases are more secured unlike the centralized transaction records maintained by a few accredited entities. This ensures that the operation of whole network is not affected by malfunctioning of a single node. Therefore, the Blockchain Technology improves the reliability of the applications (Zhang, 2016). In Higher Education it is required as certificates need to be verified often by employees and the Universities Authorities.

2. Trust: Trust is decentralized due to the use of Blockchain network. The trust bearers act as decentralized ledgers instead of relying on central governments issuing currencies, and on commercial banks. These ledgers are tamper-proofed nodes that are shared among Miners (Underwood, 2016).

3. Security: The Blockchain network security is ensured by using the forward (one-way) hash-function, to get fixed length binary sequence from a mathematical function that takes a variable length input string. The output bears no apparent relationship to the input. The process is virtually irreversible as, given just the output, the input cannot be determined (Jesse, 2016). The linear sequence of time is followed by the new block created (Salman et al, 2018).

4. Efficiency: Blockchain Technology can make pre-set procedures efficient by reducing the cost of labor and time saved. This is achieved by avoiding intermediaries to enhance the reconciliation and settlement time of the processes. The automation of distributed ledgers leads to faster settlement as achieved in digital currency of Blockchain 1.0 (Abdel, 2019).

Hence, the single PC cannot be sufficient to process so much data and it will need a pool of computing resources. This will lead to the UAE locals to secure employment with sufficient earning potential in the future. Therefore, it will serve the economic need of the Society, country. (Grech, 2017). A consortium or Government Agency like KHDA/Higher Education Ministry could manage Blockchains, where members make the decisions about how blocks are processed. In addition, Blockchains can be private where one organization controls everything. Some universities like MIT might be interested in hosting a private Blockchain or some group of universities. An ecosystem running on Open Badges, which generates digital representations of learning and skills, can use a consortium Blockchain.

10.5. How will Blockchain Transform the Higher Education System?

The popularity of the Blockchain is due to its superior cybersecurity capabilities, due to the decentralized feature and seen the increase in number of industrial applications that need security features including education, finance and healthcare. The potential uses for the Blockchain will revolutionize classroom management in the future (Ayers, 2019).

Smart classrooms are not too far off, and Blockchain Technology may become an integral part of schools all over the globe in a few years (Wiesner, 2019).

Better Security

Everyone is worried about privacy and the security of his or her data. Schools and parents are especially protective of children's data, and the threat of data breaches on online records is a serious concern.

Security Concerns and Degree Verification on College Campuses (Hafiza, 2019). Security and verification are necessary for Students trying to be employed and is a major concern for Colleges. (Salman, 2018)

Online Teaching and Learning

Blockchain Technology has been successfully used for online teaching and learning, using virtual classrooms. It can bring teachers, students, Employers, Government agencies, Administrators all on a single platform. Assessments, exams can be conducted and results recorded. Blockchain Technology used in Higher Education Domain can achieve this (Wiesner, 2019).

Library and Information Services

The Blockchain can be used to track and store information to enhance library and information services in schools. Though few libraries have started experimenting with Blockchain technology (Guang, January 2018).

Smart Contracts

Smart Contracts may be the best Blockchain application currently. The automated payments and transfer of currency or assets may work out well. Smart contract might be used to pay teacher salaries on specific dates or to make payment for equipment got from suppliers for the classroom (Cong et al, 2019).

Transformation model (Mezirow) can be applied to the change from traditional Higher Education teacher centric models to Learner centric models using modern technologies like Blockchain, Ai etc. It explains the need for the stakeholders to be accept this change, get prepared to change and change in mindset. The people concerned should stress on being inclusive, discriminating, reflective, open, and emotionally able to change (Taylor, 2017).

10.6. How Blockchain will benefit Society?

It will help with the Smart Government initiatives and is a giant jump into the development of Higher Educational needs of the community and society. It will support the society by creating Jobs in terms of Miners, coders, PC domain node maintenance for all the transactions created and to be maintained. For example, in Bitcoin, the need for mining has created more than 1 million nodes, miners along with the maintenance for these servers in the network (Ali, 2019). If we look at multi billion market across the globe, the requirements of nodes and miners will be massive. The Miners are rewarded with Tokens like Taelim Coins for creating and maintaining blocks. (Muhamud,2018) Job creation for the local population can be sufficed by a single Government initiative and investment and can take care of the job requirements in the near future (Svein,2017).

Social Model of Higher Education Benefits includes benefits like enhanced economic growth due to higher employment, higher societal production, fast technological adoption of change, and development of government and business organizations, improved well-being, women empowerment and increased social values. This leads to less crimes and wellness of the society (Behrman et al, 1994), (Williams, 2019).

11. LIMITATIONS AND FUTURE RESEARCH RECOMMENDED

The literature review is the outcome of secondary research and researching on the topic. Secondary data needs to be updated quite frequently till the paper is published which is difficult to achieve. Hence the use of the information obtained may be restricted. Moreover, secondary data might be available but may not include all the required information (Johnston, 2017). Secondary data has no control over its accuracy. Research conducted may be biased to support the vested interests of the source, known as grey material (Bornmann et al, 2019). It appears that digital transformation needs to be further explored especially in the UAE and in the Higher Education industry. Also, longitudinal survey samples can be attempted in the future. Another area of interest will be conduct focus group interviews for the Questionnaire finalization as Qualitative methodology.

The future research can make use of primary data through surveys or interviews. Limited research is carried out to correlate impact of the digital revolution within Higher Education (Almalki, 2016).

12. CONCLUSION

This paper endeavors to give an overview on the Blockchain Technology, and details on applying Blockchain Technology in Higher Education. The benefits and features of the Blockchain Technology have been reviewed and how they can be applied to Higher Education has been discussed. The limitations and gaps have been identified for future Research, to find solution so that the Blockchain Technology can be applied to its full potential and in a commercial manner to earn Maximum benefit. The contribution has been the solution for scalability (Volume of scale), the energy requirements, need for Regulation and First Sponsor Organization suggested and to be area of further Research.

For Future prospects, Research contribution can be first hand survey done with experts, Interviews to confirm the findings and including factors like energy requirements, regulator, First Sponsor Organization to get over the challenges of the Blockchain Technology to implement successfully in the Higher Education sector. The suggested Model is tested and validated using PLS method and the main contribution of building primary data for the Blockchain application in Education to be successful.

REFERENCES

1. Abdel, D. (March 2019). Using Blockchain in financial. UAE: Economic Studies Arab Monetary Fund. 139–160 (2018)
2. Ahmed, I., & Shilpi, M. A. (2018). Blockchain Technology A Literature Survey.
3. Alam, T., &Benaida, M. (2020). Blockchain and Internet of Things in Higher Education. Universal Journal of Educational Research, 8(5), 2164-2174.
4. Alexopoulos, C., Charalabidis, Y., Androutsopoulou, A., Loutsaris, M. A., &Lachana, Z. (2019, January). Benefits and obstacles of Blockchain applications in E-Government. In Proceedings of the 52nd Hawaii International Conference on System Sciences.
5. AlexrGrech, A. F. (2017). Blockchain in Education. Luxembourg: European Union 2017.

6. Ali Alammary, S. A. (13 June 2019). Blockchain-Based Applications in Education A Systematic Review. Applied Sciences — Open Access Journal, Appl. Sci. 2019, 9, 2400; doi:10.3390/app9122400.
7. Almalki, S. (2016). Integrating Quantitative and Qualitative Data in Mixed Methods Research--Challenges and Benefits. Journal of education and learning, 5(3), 288-296.
8. Alneyadi, S., Sithirasenan, E., &Muthukkumarasamy, V. (2016). A survey on data leakage prevention systems. Journal of Network and Computer Applications, 62, 137-152.
9. Ark, T. V. (2018, August 20). 20 Ways Blockchain Will Transform Education. Forbes Media LLC.
10. Ark, T. V. (2018, June 21). Imagining a Blockchain University. Forbes Media LLC.
11. Ayers, R. (JANUARY 31, 2019). How will blockchain transform the education system? Dataconomy, 2.
12. Batubara, F. R., Ubacht, J., & Janssen, M. (2018, May). Challenges of Blockchain Technology adoption for e-government: a systematic literature review. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (pp. 1-9).
13. Beck, R. (2018). Beyond Bitcoin: The Rise of Blockchain World. IEEE XPLORE.
14. Behrman, J., & Stacey, N. (Eds.). (1997). The Social Benefits of Education. Ann Arbor: University of Michigan Press. Retrieved March 15, 2020, from www.jstor.org/stable/10.3998/mpub.15129
15. Benitez, J., Henseler, J., Castillo, A., &Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. Information & Management, 57(2), 103168.
16. Booth, A., Sutton, A., &Papaioannou, D. (2016). Systematic approaches to a successful literature review. Sage.
17. Bornmann, L., Wray, K. B., &Haunschild, R. (2019). Citation concept analysis (CCA): a new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. Scientometrics, 1-24.
18. Bryson, D., Penny, D., Goldberg, D. C., & Serrao, G. (2017). Blockchain Technology for government. Montgomery, AL: The MITRE Corporation.
19. Cheah, J. H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM. International Journal of Contemporary Hospitality Management.
20. Cheah, J.H., Memon, M.A., Chuah, F., Ting, H., Ramayah, T.: Assessing reflective models in
21. Chen, G., Xu, B., Lu, M., & Chen, N. S. (2018). Exploring Blockchain Technology and its potential applications for education. Smart Learning Environments, 5(1), 1.
22. Chen, Y. (2018). Blockchain tokens and the potential democratization of entrepreneurship and innovation. Business Horizons, 61(4), 567-575.
23. Cheng, L., Liu, F., & Yao, D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(5), e1211.
24. Cheong, J. W., Muthaly, S., Kuppusamy, M., & Han, C. (2020). The study of online reviews and its relationship to online purchase intention for electronic products among the millennials in Malaysia. Asia Pacific Journal of Marketing and Logistics.
25. Cong, L. W., & He, Z. (2019). Blockchain disruption and smart contracts. The Review of Financial Studies, 32(5), 1754-1797.
26. Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.
27. Drolet, A.L. and Morrison, D.G. (2001), "Do we really need multiple-item measures in service research?", Journal of Service Research, Vol. 3 No. 3, pp. 196-204
28. Fawcett, A. (2017, March 3). Al Tamimi and Co. Retrieved from https://www.tamimi.com/law-update-articles/new-tech-on-the-block-like Dubais-Blockchain-strategy-and-why-it-matters/
29. Fink, A. (2019). Conducting research literature reviews: From the internet to paper. Sage publications.
30. Firdaus, A., Ab Razak, M. F., Feizollah, A., Hashem, I. A. T., Hazim, M., &Anuar, N. B. (2019). The rise of "Blockchain": bibliometric analysis of Blockchain study. Scientometrics, 120(3), 1289-1331.
31. Ganne, E. (2018). Can Blockchain revolutionize international trade? Geneva: World Trade Organization 2018.
32. Grover, P., & Kar, A. K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. Global Journal of Flexible Systems Management, 18(3), 203-229.

33. Guang Chen, B. X.-S. (Published: 03 January 2018). Exploring Blockchain Technology and its potential applications for education. Springer, 10.
34. HafizaYumna Email, M. M. (2019). Use of Blockchain in Education: A Systematic Literature Review. Asian Conference on Intelligent Information and Database Systems (pp. 191-202). London: Springer, Cham.
35. Hair, J.F., Ringle, C.M. and Sarstedt, M. (2011), "PLS-sem: indeed a silver bullet", The Journal of Marketing Theory and Practice, Vol. 9 No. 2, pp. 139-151.
36. Han Sun, X. W. (2018, October 10). Application of Blockchain Technology in Online Education. International Journal of Emerging Technologies in Learning (iJET) 13(10):252. doi: https://doi.org/10.3991/ijet.v13i10.9455
37. Hart, C. (2018). Doing a literature review: Releasing the research imagination. Sage.
38. Henseler, J., Hubona, G., Ray, P.A.: Using PLS path modeling in new technology research: updated guidelines. Ind. Manag. Data Syst. 116(1), 2–20 (2016). https://doi.org/10.1108/ IMDS-09-2015-0382
39. Holotescu, C. (2018). Understanding Blockchain opportunities and challenges. In Conference proceedings of» eLearning and Software for Education «(eLSE) (Vol. 4, No. 14, pp. 275-283). " Carol I" National Defence University Publishing House.
40. Huang, Z., Su, X., Zhang, Y., Shi, C., Zhang, H., &Xie, L. (2017, December). A decentralized solution for IoT data trusted exchange based-on Blockchain. In 2017 3rd IEEE International Conference on Computer and Communications (ICCC) (pp. 1180-1184). IEEE.
41. Hughes, L., Dwivedi, Y. K., Misra, S. K., Rana, N. P., Raghavan, V., &Akella, V. (2019). Blockchain research, practice and policy: Applications, benefits, limitations, emerging research themes and research agenda. International Journal of Information Management, 49, 114-129.
42. IttayEyal, E. G. (7, July 2018). Majority is not enough: bitcoin mining is vulnerable. Magazine, Communications of the ACM, Volume 61 Issue, 95-102.
43. Jesse Yli-Huumo, D. K. (October 3, 2016). Where Is Current Research on Blockchain Technology? —A Systematic Review. PLOS ONE | DOI: 10.1371/journal.pone.0163477, 1-27.
44. Johnson, K. D. (2018). BLOCKCHAIN TECHNOLOGY.
45. Johnson, R. B., & Christensen, L. (2019). Educational research: Quantitative, qualitative, and mixed approaches. SAGE Publications, Incorporated.
46. Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. Qualitative and quantitative methods in libraries, 3(3), 619-626.
47. Joseph, F.H., Black, W.C., Babin, B.J. and Anderson, R.E. (2010), Multivariate Data Analysis, 7th ed., Pearson Prentice Hall, Upper Saddle River, NJ.
48. Kem Z. K. Zhang, J. Y. (2016, December 13). Blockchain-based sharing services: What Blockchain Technology can contribute to Smart Cities. Springer open, Sun et al. Financial Innovation (2016) 2:26. doi:DOI 10.1186/s40854-016-0040-y
49. Khachaturova E. A, Makarevich M. L. (2018) Blockchain technologies: development prospects and problems of legal regulation Innovative Economy: Prospects for Development and Improvement. 2(28) 105–114
50. KHDA. (2019, September 20). Knowledge and Human Development Authority. Retrieved from KHDA: www.khda.ae
51. Koteska, B., Karafiloski, E., &Mishev, A. (2017). Blockchain implementation quality challenges: a literature. In SQAMIA 2017: 6th Workshop of Software Quality, Analysis, Monitoring, Improvement, and Applications (pp. 11-13).
52. Kraft, D. (March 2016, Volume 9, Issue 2). Difficulty control for Blockchain-based consensus systems. link. Springer, 397–413.
53. Kurt Fanning, D. P. (July 2016). Blockchain and Its Coming Impact on Financial Services. Journal of Corporate Accounting & Finance 27(5):53-57, 53-57.
54. Lakhani, M. I. (Jan 2017). The Truth about Blockchain. Haward Business Review, P: 4.
55. Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Annals of internal medicine, 151(4), W-65.
56. Lizcano, D., Lara, J. A., White, B., &Aljawarneh, S. (2019). Blockchain-based approach to create a model of trust in open and ubiquitous higher education. Journal of Computing in Higher Education, 1-26.

57. López-Cózar, E. D., Orduña-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In Springer handbook of science and technology indicators (pp. 95-127). Springer, Cham.
58. Machi, L. A., & McEvoy, B. T. (2016). The literature review: Six steps to success. Corwin Press.
59. Mahmood, Z., Arun, K. C., Rana, E., & Iftikhar, W. (2020). A Study on Issues and Challenges of Blockchain Technology in Malaysian Higher Education Institutes. International Journal of Psychosocial Rehabilitation, 24(05).
60. Malak, L. A. (2018, March 5). Like Dubai Home to the biggest Educational Blockchain Implementation Project. Retrieved from UNLOCK: https://www.unlock-bc.com/news/2019-10-04/ton-network-by-telegram-coming-late-october
61. marketing research: a comparison between PLS and PLSc estimates. Int. J. Bus. Soc. 19(1),
62. Mertler, C. A. (2018). Introduction to educational research. Sage publications.
63. Mike Sharples, J. D. (07 September 2016). The Blockchain and Kudos: A Distributed System for Educational Record, Reputation and Reward. European Conference on Technology Enhanced Learning (pp. pp 490-496). London: Springer Link.
64. Miles, M. B., Huberman, A. M., Huberman, M. A., & Huberman, M. (1994). Qualitative data analysis: An expanded sourcebook. sage.
65. MuhamedTurkanovi, M. H. (January 5, 2018). EduCTX: A Blockchain-Based Higher Education Credit Platform. IEEE, VOLUME 6, 2018.
66. Nakamoto, S. (2008, October 31). Bitcoin: A Peer-to-Peer Electronic Cash System. Retrieved from https://www.cryptovest.co.uk/: https://www.cryptovest.co.uk/
67. Nowiński, W., &Kozma, M. (2017). How can Blockchain Technology disrupt the existing business models?. Entrepreneurial Business and Economics Review, 5(3), 173-188.
68. Panda, S. S., Mohanta, B. K., Satapathy, U., Jena, D., Gountia, D., & Patra, T. K. (2019, October). Study of Blockchain Based Decentralized Consensus Algorithms. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 908-913). IEEE.
69. Quinlan, C., Babin, B., Carr, J., & Griffin, M. (2019). Business research methods. South Western Cengage.
70. Raosoft Inc (2020), Raosoft Sample Size Calculator, available at: http://www.raosoft.com/samplesize.html (accessed 1 May 2020).
71. Risius, M., &Spohrer, K. (2017). A Blockchain research framework. Business & Information Systems Engineering, 59(6), 385-409.
72. Rowley, J., & Slack, F. (2004). Conducting a literature review. Management research news.
73. Salman, T., Zolanvari, M., Erbad, A., Jain, R., &Samaka, M. (2018). Security services using Blockchains: A state of the art survey. IEEE Communications Surveys & Tutorials, 21(1), 858-880.
74. Sarstedt, M., Christian M., Ringle, D.S. and Reams, R. (2014), "Partial least squares structural equation modeling (PLS-SEM): a useful tool for family business researchers", Journal of Family Business Strategy, Vol. 5, pp. 105-115
75. Sharma, A. (2018, december 15). Hackernoon . Retrieved from Hackernoon.com: https://hackernoon.com/@AshishSharma31
76. SveinØlnes, J. U. (2017). Blockchain in government: Benefits and implications of distributed ledger technology for information sharing. Sogndal, Norway: Article in Government Information Quarterly · October 2017. doi:DOI: 10.1016/j.giq.2017.09.007
77. Swan, M. (2017). Anticipating the economic benefits of Blockchain. Technology innovation management review, 7(10), 6-13.
78. Taylor, E. W. (2017). Transformative learning theory. In Transformative learning meets bildung (pp. 17-29). Brill Sense.
79. Truby, J. (2018). Decarbonizing Bitcoin: Law and policy choices for reducing the energy consumption of Blockchain technologies and digital currencies. Energy research & social science, 44, 399-410.
80. Tschorsch, F., & Scheuermann, B. (02 March 2016). Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies. IEEE, 2084 - 2123.
81. Underwood, S. (October 2016). Blockchain beyond Bitcoin.
82. Verma, C., & Dahiya, S. (2016). ICT Adaption Model for Students: Usability & Availability, Problems & Solutions.
83. Wiesner, T. (2019, MAY 14). Blockchain Technology and the Education System. Retrieved from Blockchain for Education: https://vomtom.at/Blockchain-in-education/

84. Williams, P. (2019). Does competency-based education with blockchain signal a new mission for universities?. Journal of higher education policy and management, 41(1), 104-117.
85. Wu, B., & Chen, X. (2017). Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model. Computers in Human Behavior, 67, 221-232.
86. Xie, J., Tang, H., Huang, T., Yu, F. R., Xie, R., Liu, J., & Liu, Y. (2019). A survey of Blockchain Technology applied to smart cities: Research issues and challenges. IEEE Communications Surveys & Tutorials, 21(3), 2794-2830.
87. Zaki, I. (2019, April 30). Benefits of Blockchain Technology in the Education System. Retrieved from Moonwhale : https://moonwhale.io/Blockchain-technology-education-system/
88. Zhao, J. L., Fan, S., & Yan, J. (2016). Overview of business innovations and research opportunities in Blockchain and introduction to the special issue.
89. Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services, 14(4), 352-375.
90. Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017, June). An overview of Blockchain technology: Architecture, consensus, and future trends. In 2017 IEEE international congress on big data (BigData congress) (pp. 557-564). IEEE.

ANNEXURE 1

Questionnaire
Demographics Section: Gmail compulsory *

1. Please specify your Age group *
    a. 18-25
    b. 26-35
    c. 36-45
    d. 46-60
    e. 60 +
2. Please specify your gender *
    a. Male
    b. Female
3. Please specify your highest qualification achieved
    a. High School
    b. Undergraduate
    c. Masters
    d. Doctorate
4. Are you aware of Blockchain Application in Education?
    a. Extremely familiar- Expert in the Field
    b. Very familiar- working on Blockchain Application
    c. Somewhat familiar- only researching and yet to work on the Blockchain.
5. What has been your association with the Blockchain Technology?
    a. Researcher
    b. Student/Learner
    c. Working IT professional
    d. Business Owner
    e. Project Manager
    f. Consultant
    g. Government Official
    h. Regulator
    i. Public
    j. Trader
    k. Miner

l. Owning Domain for recording transactions

6. I believe that the Blockchain features (BCTF) that will be suitable for Education Needs are: (Express your opinion on the statement by marking the most appropriate one)

Description	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a. Security due to its working design BCTF1					
b. Transparency BCTF2					
c. Decentralized Ledgers BCTF3					
d. Minting or Corrections require approval of all concerned Parties BCTF4					
e. Immutability and Tamper deduction BCTF5					
f. Relative User Anonymity BCTF6					
g. Cost effectiveness due to faster settlement-no intermediaries BCTF7					
h. Long term Storage Ability BCTF8					

7. I believe that the needs of Education sector that drives the Business are: (Express your opinion on the statement by marking the most appropriate one)

8.

Description	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a. Confidentiality of Information HENF1					
b. Storage & Retrieval HENF2					
c. Reduced Cost HENF3					
d. Modern technologies plugins HENF4					
e. Authentication HENF5					
f. Single Regulation across the sector HENF6					
g. Quality Assurance of Education HENF7					
h. Student Centric Curriculum and Policies HENF8					

9. I believe that the main factors involved in successful implementation of Blockchain technology in Education are: (Express your opinion on the statement by marking the most appropriate one)

Description	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
a. Economies of scale SIF1					
b. Price SIF2					
c. Regulator SIF3					
d. Sponsors/Guarantor SIF4					
e. Energy needs sorted SIF5					
f. Tokenization SIF6					
g. Acceptance of BCT SIF7					
h. Behavior of Users SIF8					

10. Any other factor which you would like to recommend or something missed in this survey

---

BLOCKCHAIN ARCHITECTURE TO MEET CHALLENGES IN MANAGEMENT OF ELECTRONIC HEALTH RECORDS IN IOT BASED HEALTHCARE SYSTEMS

Maria Arif, Megha Kuliha and Sunita Varma

Department of Information Technology, S.G.S.I.T.S., Indore, M.P., India

ABSTRACT

Secure, immutable and transparent feature of blockchain has led researchers to find ways to harness its potential in sectors other than financial services. Blockchain is emerging as a popular tool to help solve some of the healthcare industry's age-old problems that have resulted in delayed treatments, inaccessible health records in emergency, wasteful spending and higher costs for doctors, health care providers, insurers and patients. Applying blockchain in healthcare brings a new challenge of integrating blockchain with Internet of Things (IoT) networks as sensor based medical and wearable devices are now used to gather information about the health of a patient and provide it to medical applications using wireless networking. This paper proposes an architecture that would provide a decentralized, secure, immutable, transparent, scalable and traceable system for management and access control of electronic health records (EHRs) through the use of consortium blockchain, smart contracts, proof-of-authentication (PoAh) consensus protocol and decentralized cloud.

KEYWORDS

Blockchain, Proof of Authentication, Smart Contracts, Internet of Things, Healthcare

1. INTRODUCTION

Safe and effective healthcare require good quality, complete, up-to-date and accurate medical records for doctors and hospital staff to make timely decisions, to improve quality of care, to develop new ways of predicting and diagnosing illness. At present, many healthcare systems still use papers and files to maintain records that often lead to delays in accessing data and hence, in providing treatment. Even where records are stored digitally as electronic health records (EHR), they mostly have server/client centralized model where server has huge pressure in terms of storage and computing, and also poses a single point of failure.

In 2019, it was reported that approximately 18 % of patient health records are duplicates and roughly one in five patients have mismatched health records, providing doctors an imperfect view of their medical history, thus resulting in delayed, improper treatment and unnecessary repeated testing. Another major concern is that though bulk of data repositories are owned by healthcare providers, pharmaceutical companies, and other stakeholders in the health and medical ecosystem, yet they do not interact with one another. This leads to non-availability of a patient's medical history to health providers in emergency cases. Sharing data between hospitals can allow for reduced costs and improved patient outcomes across hospital systems but presently, organizations and researchers cannot benefit from data sharing as patient's privacy is at stake.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.49-63, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111204

The world is witnessing an increasing number of medical records breach every year, with over 20 million breaches records in 2019 alone. The term “Medical Theft” was introduced by the World Privacy Forum (WPF) in 2006 for the illegal access and use of a patient's personally identifiable information to obtain medical treatment, services or goods. In most cases, name or health insurance numbers were used to see a doctor or get prescription drugs and in others, medical providers submitted false insurance claims for services not provided. Blockchain, if used in a well-planned architecture, can prove to be a boon to address the above issues.

Blockchain is an open, distributed, append-only public ledger technology. It consists of a chain of blocks that maintains the digitally signed transactions of the users in a verifiable and permanent way[1]. It doesn’t require the need of any central authority as the participating nodes in the network are themselves responsible for its maintenance through the use of consensus protocol which ensures that a block is added only after it has been validated by the majority of nodes. Each node in the network keeps an updated copy of the whole blockchain, ensuring consistency of the data among all nodes and protection against malicious attacks. A block once added to the blockchain cannot be altered and any changes to be made it are stored as new transactions in a new block added at the end of the blockchain, keeping the original copy intact. Hence, it ensures traceability and accountability. All the blocks are interconnected using hash values, so that any tampering with the data is easily reflected in the consequent blocks. An important feature of blockchain is the user’s anonymity which is achieved by concealing their true identity as each user is identified by their public addresses. This provides transparency in the network and allows data to be viewed and shared by all the nodes in the network.

The most common consensus protocols used today, ex. Proof of Work (PoW), have high latency in creation of a new block. Such protocols are infeasible to be used in healthcare system as IoT devices, that require fast data processing, have become an indispensable part of every healthcare system today. Others introduce some centralization, defeating the whole purpose of using a blockchain. For example, Proof of Stake (PoS) where more the number of tokens, more the power to create a block.

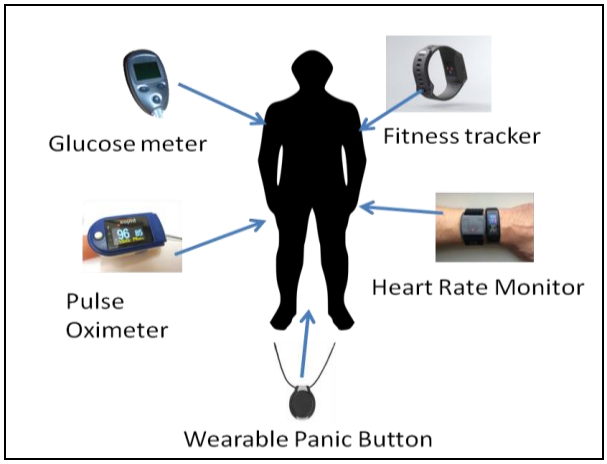


Figure 1. Role of IoT in Healthcare

The Internet of things refers to the network of physical entities that are embedded with sensors, software, and other technologies for collecting and exchanging data with other devices and systems over the internet. IoT has provided a great opportunity to build powerful industrial systems and applications by using the growing ubiquity of radio-frequency Identification (RFID) and wireless, mobile, and sensor devices [2]. Medical care and health care represent one of the

most attractive application areas for the IoT. IoT devices are used for tracking various medical issues such as electrocardiogram, blood pressure, asthma, blood sugar and so on as shown in Figure 1. The IoT has the potential to give rise to many medical applications such as remote health monitoring, fitness programs, chronic diseases, and elderly care. IoT-based healthcare services would reduce cost and increase the quality of life. But integrating the blockchain with IoT devices, which are constrained in terms of storage and computational power, brings new set of challenges.

This paper discusses these challenges and how the proposed architecture overcomes them. The remainder of this paper is organized as follows:

Section 2 discusses the related work done in this area. Section 3 discusses the structure and working of a block chain along with the benefits and challenges of integrating it with IoT in healthcare systems. Section 4 describes the architecture and workflow of the proposed model. Section 5 discusses how two blockchains were implemented based on two different consensus protocols to justify the use of proposed protocol. Section 6 presents the observed results which are discussed further in section 7. Section 8 concludes the paper by specifying the features of the proposed model that makes it fit for a secure and transparent healthcare system.

2. RELATED WORK

There have been many researches on blockchian recently. Ref. [3] surveys advances in IoT-based health care technologies and reviews the state-of-the-art network architectures/platforms, applications, and industrial trends in IoT-based health care systems. In [4] authors propose a blockchain based IoT model for medical device transactions and communication using Inter Planetary File System for storing and sharing data but they don't take into account the cost incurred by energy and time requirements of PoW protocol used for mining. Ref. [5] discuss the opportunities that blockchain offers in the field of healthcare e.g., in public health management, user-oriented medical research based on personal patient data as well as drug counterfeiting. A systematic review of the usual consensus algorithms used in the blockchain and analysis of their performance with respect to verification speed, throughput, scalability and fault tolerance has been made in [6]. In [7], authors have outlined and mapped 66 consensus protocols for private and public blockchains. The authors in [8] propose a decentralized healthcare blockchain for IoT using light weight digital signature scheme but no implementation of the same exists ensuring the low latency desired in IoT. We use the decentralized cloud model in our model inspired by them. In [9,10] authors present a novel consensus algorithm called Proof-of-Authentication (PoAh) for resource-constrained distributed systems such as the Internet of Things (IoT), edge computing and fog computing to make the blockchain application-specific. They implemented and proved that PoAh, while running in limited computer resources, has latency in the order of few seconds and is faster than PoW which is used in traditional blockchain. After considering many consensus protocols, we found PoAh to be the most apt protocol for a secure and fast data processing.

3. INTEGRATION OF BLOCKCHAIN WITH IOT

3.1. Structure and Working of Blockchain

Blockchain is a chain of interconnected blocks where each block contains data in form of multiple transactions between the nodes that are part of the network. Each block contains the transactions that occurred after the last block was added to the blockchain. The data of a particular block, when fed as input to a hashing algorithm, produces a hash that is unique to that block. This hash is also stored as a part of the block. SHA-256 is the mostly used cryptographic

hash function as it produces a 256- bit(32 bytes) one-way hash i.e. if the hash is known it is practically impossible to know the original data. And even a minute change in the input produces a totally different output hash. Thus, the hash serves as the fingerprint of the block.

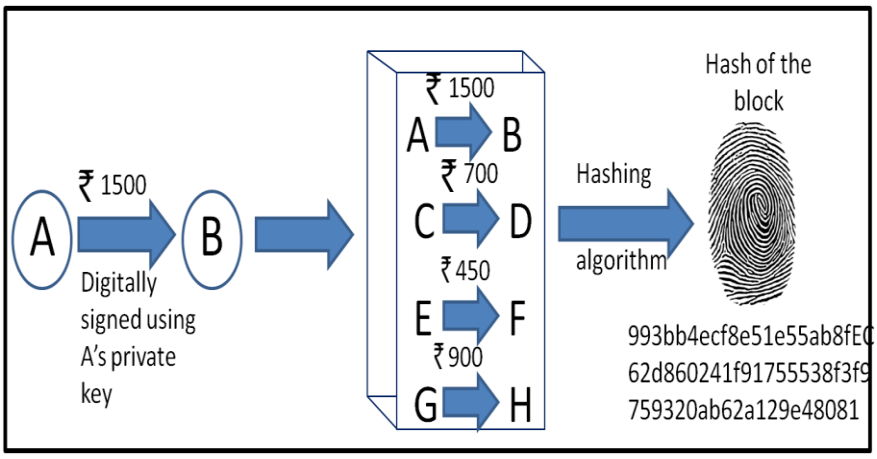


Figure 2. Pending transactions to be added to next block

Fig 2 shows how a transaction between two parties A and B is recorded in a block. Apart from data and hash, each block also contains a unique index, timestamp when it was created/ mined, a random integer nonce, and the hash of the previous block. Thus, all the blocks are connected via the hashes such that, any tampering with data in a block causes the hash to change. This renders the hash of that block stored in subsequent block, as the hash of the previous block, invalid. This provides immutability of data once stored in the blockchain and protects it against any malicious attack.

Fig 3 shows the structure of a blockchain. Each user in the network also has a unique pair of his private key (PrK) and public key (PuK). The PrK is kept as a secret whereas the PuK is known to all and also serves as the unique address of the user, hiding his true identity. The sender digitally signs (encrypts) the transaction with his PrK. The transaction is broadcasted to all the nodes in the network to be stored as pending transaction. The receivers use the PuK of the sender to authenticate the transaction. Multiple such transactions form data for the next block to be added to the blockchain.

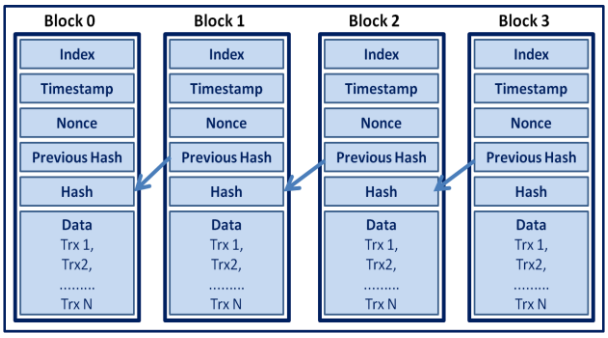


Figure 3. Structure of a blockchain

Every block needs to get consensus from majority of nodes in network before it is added to the blockchain. The process by which the nodes come at consensus at validity of the block to be added is called consensus protocol. Different such protocols exist, the most commonly used being

PoW in which all the participants of the network compete to create a block. This process is called mining, where miners (competing nodes) try to solve a computationally extensive mathematical puzzle. The puzzle is to guess a random number, nonce, such that hash obtained by sha256 (sha256 (data + nonce)) begins with a number of 0s (say seventeen, ex:00000000000000000142288235538529056 3fe84da5f0a5aa4832e8 5f68b1b5), according to the difficulty level. The only way a miner can find the number is by guessing, *i.e.* trying millions of random numbers. The one who solves the puzzle first, gets to mine (create) the block. The winner creates the block by adding the pending transactions, index, hash, timestamp, nonce and the hash of the previous block. The newly mined block is then broadcasted to all other nodes to be validated. All the nodes then check the index of the block with the one they expect, calculate and check the hash of the block, and also match the hash in the hash of the previous block field in the new block with the hash of the last block stored in their blockchain. If the block is validated by the majority of the nodes, consensus is reached and the block is added by all the nodes to their respective blockchains. This way same updated copy of blockchain is stored at all nodes, ensuring consistency. The winner who mines the block gets rewarded (with cryptocurrency in bitcoin network) for devoting its computational energy and time. If any node goes down for any reason, it broadcasts the request for the latest blockchain and stores the one which is the longest.

The consensus protocol safeguards the blockchain from malicious activity of the hacker as he may be required to devote a huge computational power and time to solve the puzzle to be able to mine the block. Apart from that he would require majority of nodes to validate his block, which would be practically infeasible for him. The protocol explained above is called Proof of Work (PoW), and is the mostly used consensus protocol in blockchains. Typically, it takes around 10 minutes for a new block to be added to the blockchain with consent from majority of the nodes.

3.2. Benefits of Blockchain Based IoT Systems

The IoT still remains in its infancy in the healthcare field due to various challenges that it imposes, the most significant being the security risk that comes with large amount of sensitive data stored in a single centralized database. Others being the issues of interoperability, scalability, flexibility, and energy efficiency [11], [12]. To address such concerns, blockchain can be a boon for IoT [13]. If IoT network is combined with the blockchain, it would provide a secure, fault tolerant, consistent healthcare system that would:

- allow storing and sharing health data securely and transparently
- enhance the accessibility of patient information in real-time
- allow secure data sharing
- ensure data integrity i.e. not changed, destroyed, or removed.
- Provide patients the control to access their data; however, they themselves won't be able to alter it either.
- ensure consistency in the patient records and increase their availability across the institutional boundaries as they may provide vital information to healthcare professionals, medical practitioners and researchers.
- guarantee medical care in emergency situations resulting in reduced suffering and medical expenses.
- aid in secure management and analysis of healthcare big data.

3.3. Challenges

Implementing blockchain with IOT may seem to be a perfect solution for healthcare systems to store highly private patient's data. But combining the two technologies brings lots of new

challenges as IOT devices have very limited storage capacity, computational power and bandwidth, whereas blockchain is computationally expensive, demands high bandwidth and storage capacity. These challenges are discussed below:

- *Scalability Issue*: IoT systems usually contain a large number of nodes. But with increasing nodes the blockchain would require more time for transaction verification and block validation.
- *Computational Capacity*: Computationally intensive Proof of Work consensus protocol in blockchain is a challenge for resource restricted IoT devices.
- *Time Consumption*: While low latency is highly desirable in most of the IoT devices that generate new data about patient's health at high frequency, mining process is highly time consuming. If this data takes too long to appear on blockchain and become available for healthcare providers, it may lead to a critical situation for the patient in emergency situations.
- *Storage Requirements*: Ever-increasing blockchain ledger has to be stored on the nodes themselves. On the contrary, IoT devices are storage constraint and usually use cloud services to extend their storage requirements. Cloud computing is based on a centralized structure whereas the whole purpose of using a blockchain is to provide a decentralized network without any central authority.
- *Access Control of Data*: Most blockchains implemented today are public/open, where anyone can join the network without the need of any permission and can access all the data stored. An important question arises here - how to provide access control to highly-classified and sensitive medical data where anyone can come in and become a part of it? Moreover, each patient may wish to share different part of his/her data with different organizations. For example, he/she may wish to make all of his/her data available to hospitals or health care providers, but only some fraction of it to insurance company or researchers.

The proposed model takes into account all the above challenges to the integrate blockchain with IoT for a healthcare network.

4. THE PROPOSED MODEL

4.1. Meeting the Challenges

To resolve the challenges discussed in the last section, the proposed model uses consortium blockchain, Proof of Authentication consensus protocol, decentralized cloud and smart contracts. Each of these is discussed in detail in this section.

4.1.1. Consortium Blockchain for Scalability

There are three main types of blockchains [14]:

- *Public/Permissionless blockchain* networks like bitcoin are completely open. These networks allow anyone to join the network without the need of any permission. Everyone in the network has full right to access all the stored data and to take part in transaction verification and consensus protocol for block validation. However, they are not the best candidate for storage and transmission of sensitive information such as healthcare records because the sole purpose of public blockchains is not to provide confidentiality but rather to allow for a publicly accessible, verifiable and unforgeable storage of data [15]. Large number of nodes taking part in consensus would result in delay in addition of blocks to blockchain.

- *Private Blockchains* are blockchains where write permissions are kept centralized to one organization/entity whereas read permissions may be public or restricted to an arbitrary extent. It is equipped with the lowest degree of openness, with a high level of access control and authority management. But the healthcare systems require openness and data sharing among multiple organizations like hospitals, researchers and insurance companies.
- C*onsortium Blockchains* are less open than the public blockchains. Only authenticated members can join the network and get access to the data recorded on the ledger. It may be apt for use in application across multiple organizations in terms of suitable degree of openness and high security. Using consortium blockchain for healthcare systems can ensure that only medical related organizations can be a part of the network and that patient's records will be in safe hands. Limiting the number of participants would result in fast transactions, privacy and high security.

4.1.2. Proof of Authentication for Time and Computational Constraint

The consensus protocol should be chosen after considering the sector requirements and deployment environment. The consensus in public environment needs to be complex and must include incentives and severe penalties for the participant nodes to ensure integrity of the network and to prevent the network from fraudulent nodes as the environment here is untrusted. Therefore, security in public blockchains is achieved at the cost of speed and scalability. On the other hand, in a private environment with trusted participating nodes, the consensus protocols can be simple and also do not require a reward mechanism as the participating nodes have business interests to protect and secure the network, therefore can focus more on speed and scalability.

The proposed model uses PoAh as a consensus protocol. Authentication uses fewer resources and less energy than other mechanisms, which can be highly advantageous in case of a resource-constrained environment like IoT architectures. PoAh utilizes minimal resources for block validation, minimal time compared to PoW without compromising security threats and it provides substantial security while integrating a blockchain based decentralized security solution to the IoT [16,17]. The working of the same is described below:

- Every participating node generates a for public-private key pair (PuK-PrK)
- There are some predetermined trusted nodes known as validators. These are initialized during the network deployment with a minimum threshold trust value, th and other network nodes with a zero trust value, tr = 0
- Network participants generate transactions with the sensed or collected data from IoT devices to form a block.
- The network users broadcast their public key, PuK, to the network and sign the block using their own private key PrK
- The nodes broadcast the blocks to the validators for validation.
- Upon receiving the block for validation, the validators authenticate the block using PuK of the sender, check the *hash of the previous block* field against the *hash* of the last block in the blockchain stored at their end and check the index expected.
- After successful authentication, validated blocks are broadcasted back to the network with the PoAh id of the validator.
- On receiving the block, the network nodes verify the PoAh information to add blocks into the chain.
- To avoid centralization of power in hands of validators, with every successful block authentication, a validator's trust value is increased by 1. Each fake block authentication decreases the trust value by 1.

- Thus, a trusted node can be out of the validation process when its trust value drops below the threshold trust value, and a normal node can be a part of the authentication process.

4.1.3. Decentralized Cloud for Storage Constraint

Most of the storage constraint IoT devices use cloud services to store the massive amount of high frequency data they generate in real time. Cloud computing is centralized as the cloud providers allow users to access the applications and computing power of their servers while also retaining complete control over those resources and their data. 80% of organizations suffered at least one cloud data breach in the past 18 months, while 43% of companies reported 10 or more cloud data breaches [18]. This may raise a question about the privacy of patient's health data. Instead of storing data directly over cloud, using blockchain at cloud level to store encrypted data may result in a decentralized cloud. The intermediate layers like patient's laptop, doctor'scomputer etc., connected with the IoT devices need not store the whole blockchain but only the hashes of the block stored at the cloud. After a block is added to the blockchain on cloud, its hash is sent back to the intermediate layer. This way any change or deletion in data would result in change of hashes which would be easily reflected at the intermediate layer, which keeps a record of hashes of all the blocks. This eliminates the need of any third party trust, because any changes in data could be easily traceable. Use of such a decentralized cloud has been proposed in this paper to overcome storage constraint and provide additional data securityas on cloud, data files are broken into fragments, encrypted and stored at multiple nodes.

4.1.4. Smart Contracts for Access Control

Access control is an essential part of the EHR and provides confidentiality by checking if a user has the required rights to access the requested resources. To assure access control of data in blockchain, which otherwise is open to all its participants, the proposed model uses smart contracts. These are lines of code (if/then statements) that are stored on a blockchain and are automatically executed when predetermined terms and conditions are met. Participants to a blockchain determine how transactions and their data are represented to other participants, agree on the rules that govern those transactions, explore all possible exceptions, and define a framework for resolving disputes. Each participant (patient, healthcare provider, etc.) defines it smart contract when it registers with the network. This includes stating which part of the data would be visible to others and also states the events to be triggered in case of exceptions. As it is too stored in the blockchain, any attempt to malicious access by tampering the smart contract would be immediately visible to all. Thus, smart contracts would provide access control without the need of any central authority.

4.2. Architecture of the Proposed Model

Figure 4 shows the architecture of the proposed model.

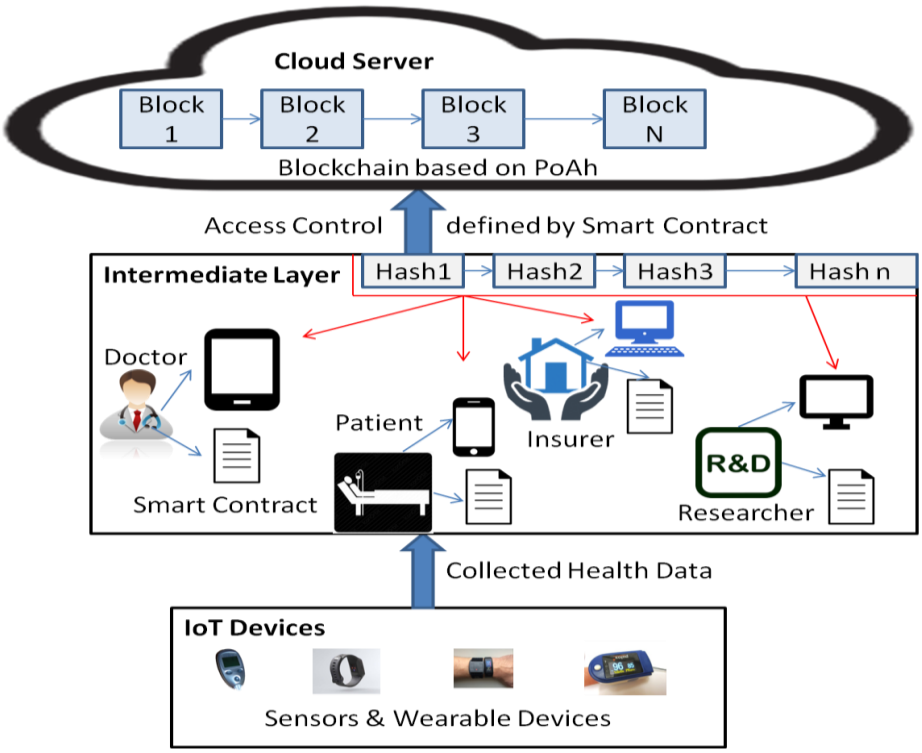


Figure 4. Architecture of the proposed model

The first layer comprises of the wearable IoT devices that collect data about the patient's health using various sensors. These may include blood pressure, heart rate or glucose monitor. The collected data is personal to the patient and can be stored on his smartphone, tablet or laptop. Each patient defines a smart contract at the time of registration. The next layer includes all the registered nodes that are the part of the network i.e. the patient, doctor, insurance provider or researcher. This layer stores only the hashes of all the blocks in the blockchain. Data from IoT devices are matched against the values specified in the smart contract and the specified event is triggered. The patient may also decide to share data with the others nodes. These actions would result in new transactions that would be broadcasted to all. The new block created would be sent and stored at the cloud. The cloud forms the next layer. This is where the whole blockchain is kept. After a new block is stored, its hash is sent back to all the nodes at the intermediate layer. The next layer comprises of the healthcare application, which is used by all the nodes to register themselves to the network, initiate various transactions and to access the blockchain data which they are authorized to. Figure 5 shows the design of the proposed model.

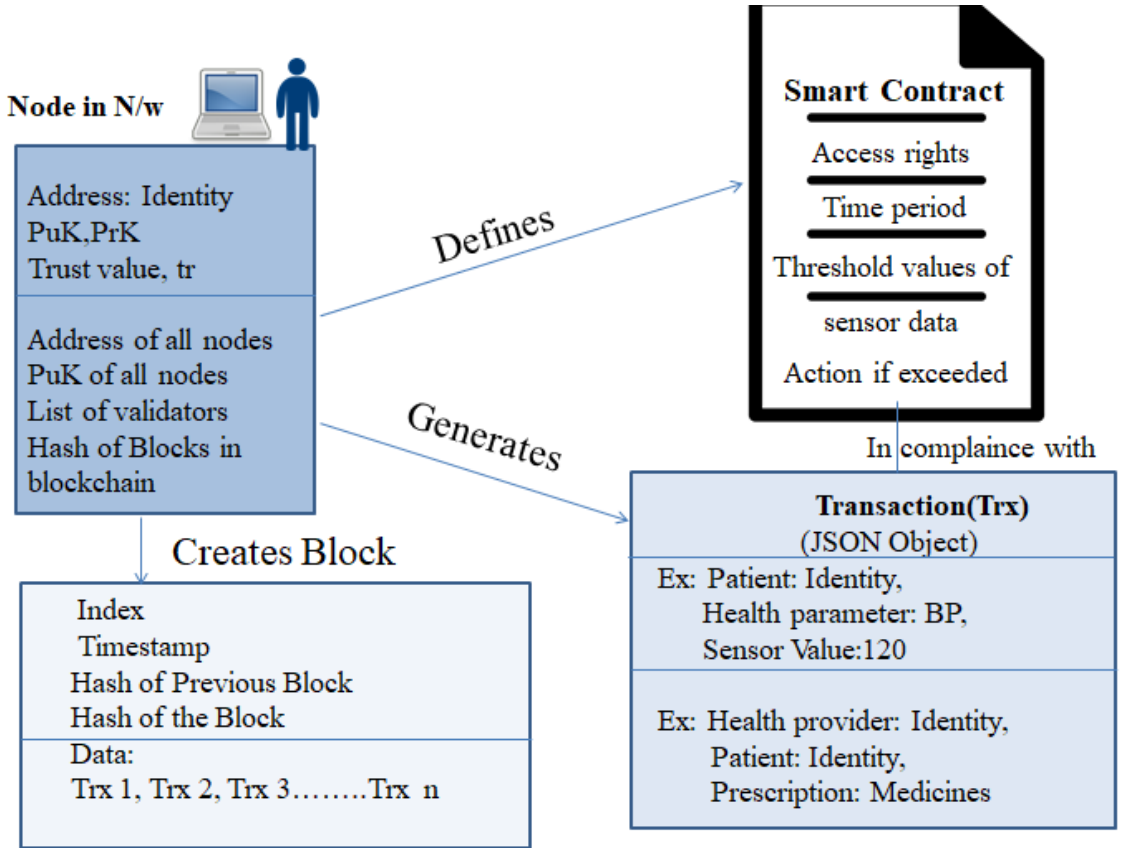


Figure 5. Design of the proposed model

4.3. Workflow of the Proposed Model

Figure 6 and 7 shows the workflow of the proposed model. It is described below in detail:

- The patient is equipped with wearable sensor devices such as a blood pressure monitor, insulin pump, temperature monitor or other known devices.
- Patients, Health providers, insurers or researchers can use the healthcare application to register to the consortium blockchain network. They can only make an account once they provide identity verification documents.

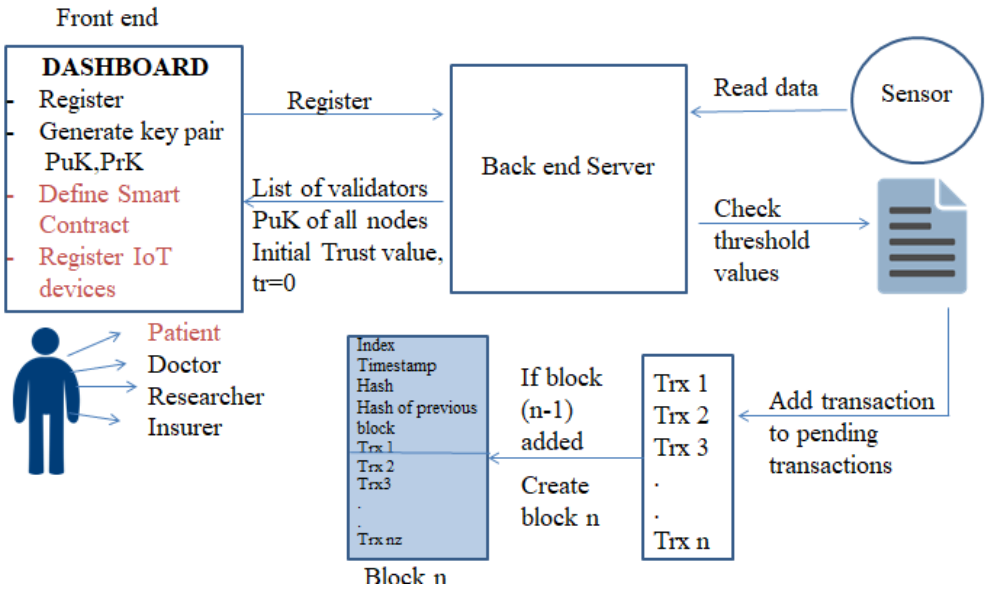
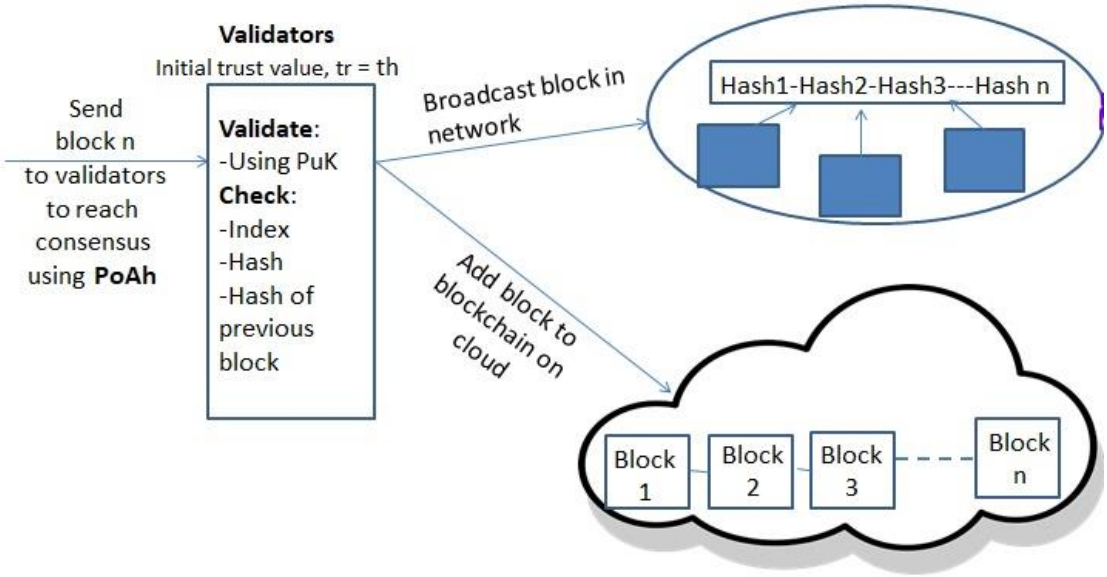


Figure 6. Workflow of block creation by a node



- **Successful validation**: tr =tr+1
- **Fake validation**: tr = tr-1
- If tr < th, Validator = Normal node
- New Validator appointed

Figure 7. Workflow (continued) of addition of block to blockchain after PoAh consensus

- Each participant creates his own private key i.e. a 256- bit number. The key is kept as a secret and a public key is generated as the hash of the private key using SHA-256 algorithm. Finally, the hash of the public key provides the address of the participant node to which all the transactions are addressed.
- Each participant has his own personal dashboard.

- The patients first define their smart contract, stating the access rights to their data and the events to be triggered in case of any violation or exceptions. The insurer and researchers can access the data according to the smart contract
- Some nodes are selected as authorized nodes and are provided an authentication id and minimum trust value. Only these validators are allowed to validate the blocks. Trust value of all other nodes is zero. All nodes in the network have information of the address of the validators.
- The health information is sent from sensors to the smart devices such as a smartphone/tablet/personal computer
- Information received is sent to the corresponding smart contract for full analysis along with the threshold values as required.
- The threshold value in the smart contract decides whether the health reading is normal as per standard readings or not.
- If the health reading is abnormal, then the smart contract would execute specified action and send an alert to the health providers in intermediate layer.
- The patient or Health provider may initiate a transaction to get treatment, pay fees or send prescription. Each transaction is signed with the private key so that it can be verified at the other end using the corresponding public key.
- All such transactions are broadcasted to all the nodes. One of the authorized nodes, creates a block with all the pending transactions and sends it to other authorized nodes with its authorization id. The block is validated by all other authorized nodes.
- Once validated the block is sent to the cloud server for storage, where it is added to the blockchain.
- The hash of the recently added block is sent back by the cloud to all nodes in the intermediate layer. Each node at this layer keeps a chronological list of hashes of all the blocks stored in blockchain on the cloud server.

5. IMPLEMENTATION OF BLOCKCHAIN

The consensus protocol is the backbone of any blockchain. To prove that PoAh is a faster and more efficient consensus protocol than the most commonly used Proof-of-Work, for the proposed architecture, two healthcare blockchain models were implemented. One used PoW as the consensus algorithm and the other PoAh. Both the models were run on the same machine, with Windows 10 operating system and 8 GB RAM, one at a time. Transactions for both were simulated using the Postman simulator. The post calls were made through Postman to simulate the registration of block and creation and broadcasting of transaction. Node js and Visual Code editor was used to write the code and to run different nodes at different ports. Node js provides many features and is very popular for javascript programs with its rich built-in libraries. HTML, CSS, Javascript, Jquery and Angular js were used to design the front end where user can register and view the data of the blockchain.

First, 15 nodes were simulated for both the models. 50 transactions were simulated to be added to the block to be created. The time taken for execution of the consensus protocol and addition of a newly created block to the blockchain was noted for the blockchains. Then, 25 nodes were simulated for both the models. 50 transactions were simulated to be added to the new block. The time was recorded in this too case. Lastly, 50 nodes were simulated and time was again noted for the block addition to the blockchain after the consensus. On an attempt to create note for further testing, a warning was shown on the system depicting overuse of resources as all the nodes were running on the same machine and PoW protocol uses a lot of computational power and memory to solve the mathematical puzzle. More computers were not available and labs were not

accessible due to the covid 19 pandemic lockdown. Thus, observations were made and results were recorded to make a comparison between the PoW and PoAh based blockchains.

6. RESULT

The two blockchain models, based on PoW and PoAh were successfully executed. Different number of nodes were simulated each time with fifty transactions. the time taken for execution of the consensus protocol in both the cases with different number of nodes was recorded. The results are produced in the table1 given below.

Table 1 Time taken to execute PoW and PoAh protocols

Number of Nodes	No. of Transactions	Time Taken in sec (PoW)	Time Taken in sec (PoAh)
15	50	29.252	0.656
25	50	67.313	1.503
50	50	888.777	2.583

7. DISCUSSION

The results clearly show that PoAh consensus protocol is faster than PoW with respect to validation and addition of a newly created block to the blockchain. With less number of nodes, it is approximately 45 times faster than PoW protocol. With the increase in the number of nodes, PoW takes more time which is approximately 14 min to mine a block whereas PoAh takes few seconds to do so. Thus, while the latency of PoW increases with the expansion of the network, PoAh remains low latent and is approximately 300 times faster. This proves that PoAh is highly scalable and can update the blockchain with the new blocks at a faster rate. Hence, PoAh would be an apt protocol to be used in a healthcare blockchain that requires low latency in addition of a newly created block to the existing blockchain.

8. CONCLUSION

An efficient model for access control and secure management of EHRs has been proposed and described in detail in this paper. The use of PoAh protocol for consensus has been justified by implementation and the observed results. Using the blockchain technology with IoT networks along with proof of authentication protocol, decentralized cloud and smart contract, would allow tamper proof medical data storage, quick reporting, data sharing and lowering the cost of medical services. Such a healthcare system is the need of the hour, especially in this pandemic of covid-19.

REFERENCES

[1] Du Mingxiao, Ma Xiaofeng, Zhang Zhe, Wang Xiangwei, Chen Qijun," A Review on Consensus Algorithm of Blockchain", 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) Banff Center, Banff, Canada, October 5-8, 2017

[2] Puthal, D., Malik, N., Mohanty, S. P., Kougianos, E., & Das, G. (2018). Everything You Wanted to Know About the Blockchain: Its Promise, Components, Processes, and Problems. IEEE Consumer Electronics Magazine, 7(4), 6–14. doi:10.1109/mce.2018.2816299

[3] Li Da Xu, Senior Member, IEEE, Wu He, and Shancang," Internet of Things in Industries: A Survey", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 4, NOVEMBER 2014

[4] Tushar Dey , Shweta Sunderkrishnan , Shaurya Jaiswal and Prof. Neha Katre ,"HealthSense: A Medical Use Case of Internet of Things and Blockchain" in Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9

[5] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for IoT security and privacy: The case study of a smart home," in Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2017, pp. 618–623.

[6] Matthias Mettler," Blockchain Technology in Healthcare The Revolution Starts Here", 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)

[7] Wan, S., Li, M., Liu, G., & Wang, C. (2019), "Recent advances in consensus protocols for blockchain: a survey. Wireless Networks". doi:10.1007/s11276-019-02195-0

[8] Ashutosh Dhar Dwivedi, Gautam Srivastava, Shalini Dhar and Rajani Singh ,"A Decentralized Privacy-Preserving Healthcare Blockchain for IoT", MDPI Article, 15 January 2019

[9] Deepak Puthal, Saraju P. Mohanty, Priyadarsi Nanda, Elias Kougianos, and Gautam Das," Proof-of-Authentication for Scalable Blockchain in Resource-Constrained Distributed Systems", 2019 IEEE International Conference on Consumer Electronics(ICCE), doi:10.1109/icce.2019.8662009

[10] Deepak Puthal, Saraju P. Mohanty, Venkata P. Yanambaka, Elias Kougianos "PoAh: A Novel Consensus Algorithm for Fast Scalable Private Blockchain for Large-scale IoT Frameworks"

[11] Riazul Islam, S. M., Daehan Kwak, Humaun Kabir, M., Hossain, M., & Kyung-Sup Kwak. (2015). "The Internet of Things for Health Care: A Comprehensive Survey", IEEE Access, 3, 678–708.

[12] Stephanie Baker, Wei Xiang, Senior Member, IEEE, and Ian Atkinson," Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities", IEEE Access, 5, 26521–26544. doi:10.1109/access.2017.2775180

[13] Yaqoob, I., Ahmed, E., Hashem, I. A. T., Ahmed, A. I. A., Gani, A., Imran, M., & Guizani, M. (2017). Internet of Things Architecture: Recent Advances, Taxonomy, Requirements, and Open Challenges. IEEE Wireless Communications", 24(3), 10–16.

[14] Kumar, N. M., & Mallick, P. K. (2018), "Blockchain technology for security issues and challenges in IoT.", Procedia Computer Science, 132, 1815–1823. doi:10.1016/j.procs.2018.05.140

[15] Alhadhrami, Z., Alghfeli, S., Alghfeli, M., Abedlla, J. A., & Shuaib, K. (2017). "Introducing blockchains for healthcare." 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA). doi:10.1109/icecta.2017.8252043

[16] Shahaab, A., Lidgey, B., Hewage, C., & Khan, I. (2019). "Applicability and Appropriateness of Distributed Ledgers Consensus Protocols in Public and Private Sectors: A Systematic Review". IEEE Access, 7, 43622–43636. doi:10.1109/access.2019.2904181

[17] Maitra, S., Yanambaka, V. P., Abdelgawad, A., Puthal, D., & Yelamarthi, K. (2020), "Proof-of-Authentication Consensus Algorithm: Blockchain-based IoT Implementation", 2020 IEEE 6th World Forum on Internet of Things (WF-IoT). doi:10.1109/wf-iot48130.2020.9221187

[18] "80% of Organizations Suffered a Cloud Data Breach in the Past 18 Months" article by CISOMAG, June 4,2020

AUTHORS

Maria Arif Pursuing Masters of Engineering in Information Technology from Shri G.S. Institute of Technology and Science, Indore (M.P.), she has an experience as a software developer in Impetus Infotech. Her fields of interest include IoT networks, Blockchain and full stack web development.

Megha Kuliha Working as Senior Assistant Professor in Information Technology Department of Shri G.S. Institute of Technology & Science, Indore, he has 14 Years of experience in academics. She is currently pursuing her PhD from RGPV Bhopal. Her research areas are Blockchain, Network Security & Cloud Computing.

Sunita Varma Born in 1969 at Indore in Madhya Pradesh, Dr. Sunita obtained B.E. (Electronics & Telecommunication, 1991) and M.E. (Computer Engineering, 1998) from Shri. G.S. Institute of Technology and Science, Indore. She had been awarded Doctoral degree from Devi Ahilya University, Indore in 2013.At present she is working as Professor and head in the Department of Information Technology at Shri. G.S. Institute of Technology and science, Indore. Her fields of interest are Mobile Computing and Communication, Cloud Computing, Big Data etc. She is professional member of several international bodies like IEEE and life member of Institute of Engineers.

MUSIC SIGNAL ANALYSIS: REGRESSION ANALYSIS

V. N. Aditya Datta Chivukula and Sri Keshava Reddy Adupala

Department of Computer science and Engineering, International Institute of Information Technology, Bhubaneswar, Odisha, India

ABSTRACT

Machine learning techniques have become a vital part of every ongoing research in technical areas. In recent times the world has witnessed many beautiful applications of machine learning in a practical sense which amaze us in every aspect. This paper is all about whether we should always rely on deep learning techniques or is it really possible to overcome the performance of simple deep learning algorithms by simple statistical machine learning algorithms by understanding the application and processing the data so that it can help in increasing the performance of the algorithm by a notable amount. The paper mentions the importance of data pre-processing than that of the selection of the algorithm. It discusses the functions involving trigonometric, logarithmic, and exponential terms and also talks about functions that are purely trigonometric. Finally, we discuss regression analysis on music signals.

KEYWORDS

Machine Learning, Regression, Music Signal Analysis.

1. INTRODUCTION

Regression analysis gained its importance when several statisticians found out its applications in the real-world such as predicting the price of land in a certain city, estimating the complex polynomials through working on the dataset provided, estimating whether a given medicine will work on a large amount of people, etc. [1] also gained its profound importance during the past decade with its description of solving various statistical models. [2] also came into the picture showing its influence over dealing with trigonometric functions, but, there are some areas where we need to understand the importance and need for a perfect combination of above-mentioned approaches in a simple way to enhance the accuracy of results and to understand the true efficiency of regression analysis in many other fields which recently growing with respect to the growing demand for new applications in research. Some primary variations of regression are [3-5], etc. These algorithms have their own importance individually and are application-specific. Therefore, the practical realization of technical research applications needs their respective algorithms or approaches which can better the efficiency and accuracy of the applications with the least error possible.

1.1. Motivation

This paper discusses about trigonometric regression and polynomial regression on hypothesis involving logarithmic or exponential terms to establish the importance of adding features to the dataset for better results. Thus, the paper also provides the contrast between the performance delivered by the above-mentioned methods and simple neural networks. Hence, by establishing

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.65-75, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111205

the context, music signal analysis is performed considering the same idea. The idea of the paper is that a proper data pre-processing step can highly reduce the error and allows us to solve problems with much more light-weight and basic methods.

2. REGRESSION ANALYSIS OF THE TRIGONOMETRIC FUNCTION

In this section we will consider a trigonometric function as shown in equation 1. To take a completely random function, we considered generating a random function. To generate a random trigonometric function, we have used the python code as provided in listing1. In the code, there is a feature list containing all features of interest. There is a single 'For' loop ranging from 0 to length of feature list. An individual is allowed to choose a range which is equal to the number of terms that are required in the end polynomial. For each iteration of the loop, we randomly select coefficient for each term and the term itself from the feature list. Then, we multiply the coefficient and store the resulting string in a list known as function. We continue the same until the loop is completed. Hence, we end up having a list of terms as strings. Finally, we join all the strings using 'join' function which results in a random trigonometric polynomial in string datatype. It should be noted that range of loop is the number of terms one desires in the end function. One other point is that, feature 'x' is not considered while generating the function as the interest of this section was to discuss a pure trigonometric function. In this section

Listing 1: Python Code for generating function with only trigonometric terms

```
feature = ['x','np.sin(x)','np.cos(x)', 'np.sin(x)*np.cos(x)']
function = []
for i in range(len(feature)):
  coef = str(np.random.choice(np.ara-
                    -nge(100)))
  term = coef + '*' + np.random.cho-
       -ice(feature[1:])
  function.append(term)
function = '+'.join(function)
function = 'y = ' + function
```

Equation (1) is the function taken to explain the importance of trigonometric features in regression analysis for this section.

$$95*np.sin(x)*np.cos(x)+37*np.sin(x)+90*np.sin(x)*np.cos(x)+45*np.sin(x)*np.cos(x) \quad (1)$$

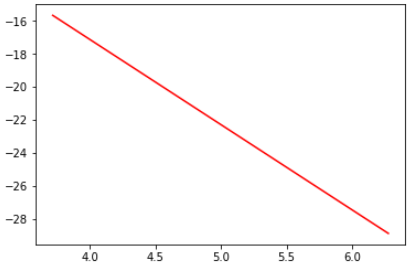


Figure 1 : plot showing predictions on y-axis with inputs on x-axis for simple linear regression

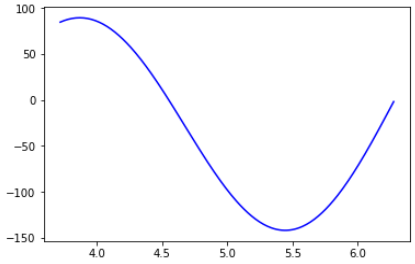


Figure 2: plot depicting desired outputs for the inputs.

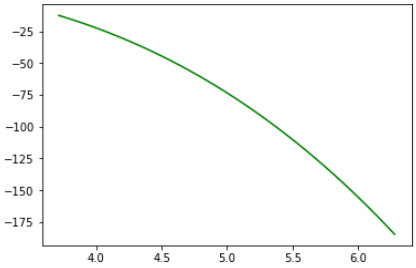


Figure 3 : plot with predictions on y-axis and inputs on x-axis for polynomial regression

If we carefully observe there are no terms with ‘x’ raised to a certain power. When we apply linear regression analysis on the dataset with input as ‘x’, where ‘x’ belongs to the range $[-\pi, \pi]$ in steps of 0.01, and output ‘y’ calculated according to the equation (1) for thousand samples, we get the graph shown in figure 1 which depicts the performance of the linear regressor on the test set, whereas the expected performance or the desired performance is as shown in figure 2. Hence, we can decide that the linear regressor performed poorly as expected. Now, if we use a polynomial regressor and consider the hypothesis degree to be 2 and train on the same training data and test it, we obtain performance as shown in figure 3. It is expected that the polynomial regressor cannot predict the trigonometric terms as there is no feature which is trigonometric in nature. Now, one can always think about using a simple neural network [6], but that also would not work as the training set is too low for the neural network to generalize the trigonometric hypothesis and training the network excessively for a greater number of epochs would result in overfitting of data and also does not assure accuracy. We can also try with [7] but, we should not forget the fact that LSTM networks require

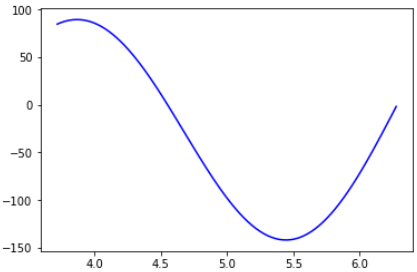


Figure 4: plot with inputs on x-axis and predictions by simple linear regression after adding trigonometric features to the dataset on y-axis.

a high amount of data and moreover are computationally expensive as compared to the simple neural networks and regression analysis discussed above. Now, if we closely look at the situation and introduce the trigonometric terms in the hypothesis considered in the case of simple linear

regression as redefined according to equation (2) and train on the dataset with a new hypothesis and apply linear regression analysis then we can observe the performance as shown in figure 4. Thus, by looking at figure 2 and figure 4, we can understand the importance of trigonometric features in linear regression provided the dataset has a trigonometric relationship. We can even look at table 1 to just checkup on the errors obtained with each regression approach discussed. Generally, trigonometric regression analysis need can be observed in the fields like signal processing and wave analysis. We are going to continue this idea as polynomial trigonometric regression in section 3 which actually makes us think to consider adding trigonometric features as a primary data preprocessing step whenever we encounter with regression analysis problems.

Table 1 Error table for pure trigonometric function by different algorithmic approaches.

ALGORITHM	ABSOLUTE ERROR
Proposed Approach	6.610267888618182e-12
Linear Regression	18573.351509906905
Polynomial Regression	15689.82990204867

3. REGRESSION ANALYSIS OF POLYNOMIAL WITH TRIGONOMETRIC FEATURES

In section 2 we have discussed function having only trigonometric terms without the mixture of linear or quadratic terms in ‘x’, where ‘x’ is the input value. Consider a function as described by equation 2 in which we observe terms such as ‘x*cos(x)’ and so on, which is difficult for simple neural networks and even the simple statistical regression algorithms like linear regression and polynomial regression to learn on minimal data.

Equation 2 is generated using the code provided by listing 2. To briefly explain the algorithm, in first loop the degree of the polynomial is kept as range and all orders of input feature 'x' are included in the features list. Then, every term in the 'terms' list is included in the features list. Now, when the 'features' list is ready, a 'function' is defined, in which, an empty list 'T' is considered and the number of terms in the generated polynomial is decided at random by keeping a maximum upper-limit. Now, a loop is considered keeping number of terms as range and for each iteration a term is appended to list 'T' by generating the term with a randomly selected number of features. Finally, polynomial is created by joining the terms stored in list 'T'.

Listing 2: Python code to generate a random mixed polynomial

```
x = np.pi # buffer value
functions = []
terms = ['np.cos(x)','np.sin(x)','np.tan(x)', 'np.log(x)','np.exp(x)']
features = []
for i in range(2):
    features.append("x**"+str(i+1))
for i in terms:
   features.append(i)
# generating function
def function():
   T = []
   number_terms = np.random.cho-
                -ice(np.arange(10))+1
   for i in range(number_terms):
```

```
    num_features = np.random.cho-
            -ice(len(features))+1
    l = []
    for j in range(num_features):
       l.append(features[np.random.cho-
       -ice(np.arange(len(features)))])
    t = '*'.join(l)
    T.append(t)
  func = '+'.join(T)
  func = 'y = ' + func
  return func
```

Hence, here too we can add the additional features which include trigonometric, logarithmic and exponential features in 'x' and also consider all permutations possible once the individual estimates the degree of polynomial the learning hypothesis would belong to in the same way as we do in case of normal polynomial regression. If we carefully observe figure 5 which depicts the predictions by support vector regression trained on dataset with inputs ranging from $-\pi$ to π and outputs calculated according to equation 2, we see that the expected plot as in Figure 6 is completely different from what has been predicted which leads to high absolute error on test set. When we apply polynomial regression analysis keeping the degree as 2, then also we can see that the plot by polynomial regression as depicted in figure 7 is mostly off in predicting the desired outputs as shown in figure 6.

$$Y=[e^{x}*\cos(x)*\tan^{2}(x)]+[x^{3}*\sin(x)]+[x^{3}*\tan(x)*\sin(x)*\log(x)]+[x^{2}]+[x^{3}*\cos(x)*\tan(x)*e^{x}*\log(x)] +[e^{x}*\tan(x)*x^{4}] \quad (2)$$

Hence, if we are able to actually consider the list of additional features which are all possible permutations of 'x' with trigonometric, logarithmic and exponential functions acting upon it and then apply linear regression analysis, we observe the desired plot as in figure 8 which is almost similar to actual relationship showcased in equation 2. If we compare figure 6 and figure 8, we can understand that the simple addition of all combination of functional features can affect the performance of an algorithm by a great extent. Table 2 depicts the errors obtained by discussed algorithms.

Table 2 Error table for polynomial with complex terms by different algorithmic approaches

Algorithm	Absolute Error
proposed approach	27.97901221743491
Support Vector Regression	14177902477532.947
Polynomial Regression	15.715957e+12

If one thinks that the number of permutations is increasing with degree of the hypothesis then he can apply dimensionality reduction techniques such as principal component analysis and thereby decreasing the computational time taken. This approach is only successful when the input is related to output with assumed combinations of features, in this case which are trigonometric, logarithmic and exponential. We can also analyze data in preprocessing stage to identify more complex functions as features in 'x' depending upon the dataset.

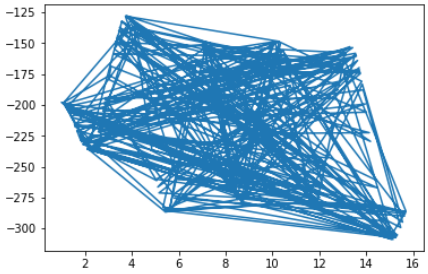


Figure 5: plot depicting predictions on y-axis and input value on x-axis by support vector regression

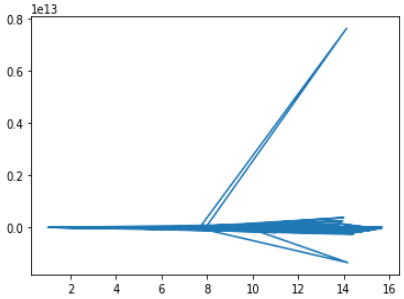


Figure 6: plot depicting expected outputs on y-axis for inputs on x-axis

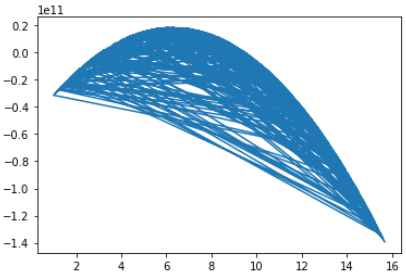


Figure 7: plot depicting predictions on y-axis for inputs on x-axis by polynomial regression

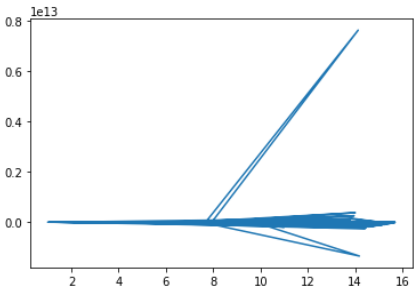


Figure 8: plot depicting predictions on y-axis for inputs on x-axis by linear regression after addition of features discussed in section 3

4. MUSIC SIGNAL ANALYSIS

Music signal is one of the complicated signals out there and definitely making an machine learning algorithm to learn from it and make it figure out parameters such as amplitude,

frequency and phase is a difficult task as the superposition of several sinusoidal waves change after very short amount of time over complete time interval, but, if we assume that there are only a constant number of waves superposed over each short time frame and consider a superposition as shown in equation 3, then we can optimize the parameters using many optimization algorithms out there. In this case we have taken gradient descent algorithm to optimize which is simple to understand and apply. Here, we considered a random background music track [8] for explanatory purpose and considered first 800,000 samples of the audio amplitudes from left channel, then, we have further divided the entire training set into 800 segments with each containing 1000 samples. These 1000 samples are trained thereby, optimizing the parameters in hypothesis which are amplitude, frequency and phase of each of the constant number of waves considered, here we assumed the constant value to be 20 for explanatory purpose. This summarizes the problem as to optimize the parameters frequency, amplitude and phase of each of the 20 waves in that particular time frame of 1000 samples using gradient descent assuming the step size as 1 and considering squared error as loss function.

$$y = \sum_{i=0}^{20} a\_i * sin(2\pi(f\_i * x + phase\_i)) \quad (3)$$

a_i – amplitude parameter of i^{th} wave
f_i - frequency parameter of i^{th} wave
phase_i – phase parameter of i^{th} wave

One can always experiment upon different optimizing algorithms and consider different values for the hyperparameters mentioned according to the audio data they have. We have also normalized the time frame values which act as input by dividing each value on time axis with 44100 and then subtracting the mean from the input array and finally dividing it with the standard deviation. Two approaches have been followed to actually perform regression analysis as described above. The first approach is simple way of optimizing all the parameters of a particular time frame simultaneously at each step of gradient descent [9], but, this method forces the waves to learn independently of each other which results in same optimized parameters for each wave, that is, for example if frequency is 1, amplitude is 1 and phase is 0 for the first wave in the hypothesis after optimizing, then, the each of the remaining 19 waves of that time frame will also have the same values for frequency, amplitude and phase respectively. From first approach one can easily understand that the conventional form of regression analysis cannot be performed for music signal and hence, we have considered a second approach which is to optimize the second wave with respect to first, third with respect to second and first, and so on, similar to cost functions described by Algorithm 1.

As shown in the algorithm 1 we can update array ‘h’ which stores the superposition values of all ‘i’ number of waves when optimizing ‘i+1’ wave’s parameters, so that, the superposition value can be added to redefine cost function for each wave pertaining to the same time frame and thereby, optimizing the parameters of each wave with respect to the values obtained by the superposition of previous waves. The figure 9 represents the plot between the desired amplitudes and the time, and, figure 10 shows the plot obtained by the hypothesis considered, which is the superposition of 20 sine waves, but, the plot as in figure 10 is obtained by calculating amplitudes according to equation 4, where we did not consider amplitude parameter of each sine wave of that time frame as they were not even close to the desired values and scaling up the error by large extent which can be observed in figure 11. This is a drawback currently but, if followed a different technique of optimization for amplitude parameter, then definitely we can make this approach work.

$$y = \sum_{i=0}^{20} sin(2\pi(f\_i * x + phase\_i)) \quad (4)$$

One important thing is that, for optimizing amplitude or frequency or phase we considered the gradients as follows:

$$Ga = step * (h + a\_i * sin(2\pi x) - y) * (sin(2\pi x)) \quad (5)$$
$$Gf = step * (h + sin(2\pi * f\_i * x) - y) * (2\pi x * cos(2\pi * f\_i * x)) \quad (6)$$
$$Gp = step * (h + sin(2\pi x + 2\pi p\_i) - y) * (2\pi * cos(2\pi x + 2\pi p\_i)) \quad (7)$$

Ga – amplitude gradient, Gf – frequency gradient, Gp – phase gradient

---

Algorithm 1: Optimization

Input: data x, size n; amplitudes y, size n; step s; starting index of time frame start
Input: parameters parameters, size (20, 3)

Initialize h = array(zeros(1000))

for k=0 to 19 do
for j=0 to 9 do
for i=start to start+1000 do
Assign Ga=step*(h_i+ parameter s_{k0}*(sin(2πx_i)-y_i)*(sin(2π*x_i))
Assign
Gf = step*(h_i+sin(2π*$parameters_{k1}$*x_i)-y_i)*(2π*$x_i$cos(2π*$parameters_{k1}$*x_1))
Assign
Gp = step*(h_i+sin(2π*x_i+2π*$parameters_{k2}$)-y_i)*(2πcos(2πi*x_i+2π*$parameters_{k2}$))
Assign $parameters_{k\_0}$ = $parameters_{k\_0}$-Ga
\STATE Assign $parameters_{k\_1}$ = $parameters_{k\_1}$-Gf
\STATE Assign $parameters_{k\_2}$ = $parameters_{k\_2}$-Gp
\ENDFOR
\ENDFOR
for v=start to start+999 do
Assign w = vmod1000
Assign
h_w = h_w + ($parameters_{k0}$ * np.sin(2 * np.π * ($parameters_{k1}$ * x_v + $parameters_{k2}$)))
ENDFOR
ENDFOR

Where, we only consider the effect of the parameter for which we compute the gradient. For example, while computing the gradient for amplitude parameter we make f_i as 1 and p_i as 0 and hence, we are optimizing only amplitude with respect to the samples, which is to try fit amplitude parameter for that wave for that time frame completely. Similar pattern can be observed for frequency where a_i is made 1 and p_i as 0 and in case of phase gradient a_i is 1 and f_i is also one. This can be understood as independent parameter training for which we got the results as shown in figure 10. We have also considered dependent parameter training where we try to optimize one with respect to other, for which the gradients are as follows:

$$Ga = step * (h + a\_i * sin(2\pi * f\_i * x) - y) * (sin(2\pi * f\_i * x)) \quad (8)$$
$$Gf = step * (h + sin(2\pi * f\_i * x) - y) * (2\pi x * cos(2\pi * f\_i * x)) \quad (9)$$
$$Gp = step * (h + a\_i * sin(2\pi * f\_i * x + 2\pi p\_i) - y) * (2\pi * a\_i * cos(2\pi * f\_i * x + 2\pi p\_i)) \quad (10)$$

Ga – amplitude gradient, Gf – frequency gradient, Gp – phase gradient

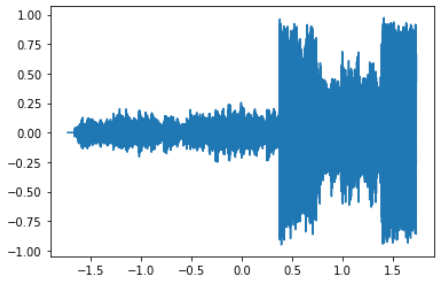


Figure 9: plot original audio data with desired amplitudes on y-axis and time period on x-axis.

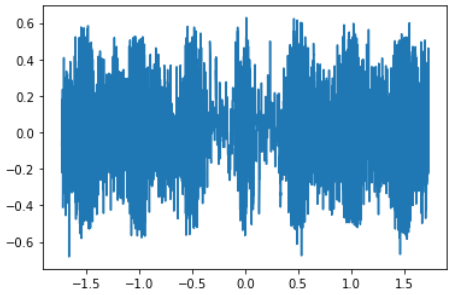


Figure 10: plot with predicted amplitudes on y-axis and time period on x-axis by following independent parameter training excluding amplitude parameter.

where, frequency is computed independently and amplitude is computed with respect to frequency parameter and finally phase parameter is computed with respect to frequency and amplitude parameter. For, dependent parameter training we observed a higher loss and hence, currently independent parameter training is better. On an important note, as we have not predicted the amplitude parameter for 20 waves of each time frame properly, we have divided the final value by 20 which should be the mean amplitude at that particular instant. The plot observed for dependent parameter training can be observed in figure 12 and we can also observe figure 11 in which we calculated amplitudes considering amplitude parameter for each of 20 waves in that time frame and clearly decide why we did not consider amplitude parameter.

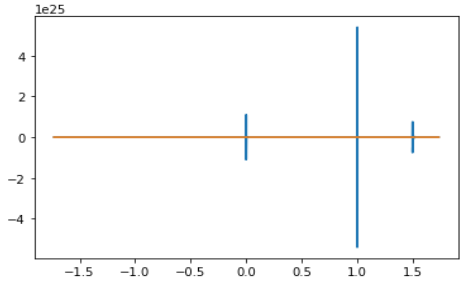


Figure 11: plot with predicted amplitudes on y-axis and time period on x-axis by following independent parameter training including amplitude parameter where horizontal plot represents original signal.

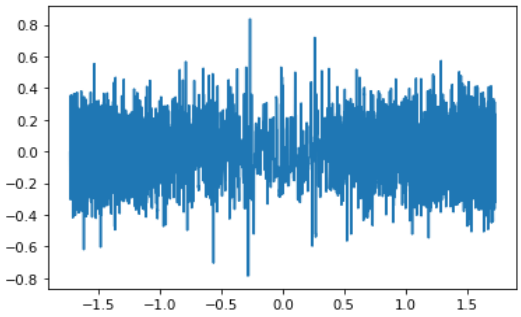


Figure 12: plot with predicted amplitudes on y-axis and time period on x-axis by following dependent parameter training.

5. CONCLUSION

Regression algorithm is the most fundamental and important algorithm which can be powerful when hypothesis, optimization and features are selected properly. It has the potential to even perform better than the current advanced machine learning techniques. With this theory we try to propose that, as algorithm selection is important for an application, similarly, data pre-processing and hypothesis reformulation is also that important. We need to focus on formulating the underlying functions in pre-processing stage itself so that even on less amount of data, the algorithm can perform much more efficiently and we can eliminate the risks such as underfitting or overfitting. This also specifies that we need to conduct more experiments with each algorithm by reformulating some of its parts on the dataset, so that, we can understand some of the relationships in the dataset and even have a combination of different machine learning algorithms acting on same dataset which may be much more efficient, and also understand the power of interdisciplinary algorithms. This also sheds light on the fact that adding features by exploring dataset can boost algorithm's performance and efficiency.

ACKNOWLEDGEMENTS

The authors would like to thank Abhiram Reddy for his participation, assistance and his valuable time. We would like to thank our guide, Mr. Rupaj Kumar Nayak for his guidance for this work.

REFERENCES

[1] Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. The Annals of Statistics, 2873-2903.
[2] Eubank, R. L., & Speckman, P. (1990). Curve fitting by polynomial-trigonometric regression. Biometrika, 77(1), 1-9.
[3] Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. Radiology, 227(3), 617-628.
[4] Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.
[5] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
[6] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.
[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[8] DJ Sakthi. (2020, April 21). Charlie Bgm Mix[Video]. YouTube. https://www.youtube.com/watch?v=nopQ6TT_pGo
[9] Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

AUTHORS

V. N. Aditya Datta Chivukula is currently an undergraduate student in the Department of Computer Science and Engineering at International Institute of Information Technology, India. His area of interests are in Machine learning, Deep learning and Natural Language Processing.

Sri Keshava Reddy Adupala is currently an undergraduate student in the Department of Computer Science and Engineering at International Institute of Information Technology, India. His area of interests are in Data Analytics, Data Visualization and Machine Learning.

CLASSIFICATION OF MAMMOGRAPHIC IMAGES BY *OPENVINO*: A PROPOSAL OF USE TO ENHANCE MORE EFFECTIVITY IN CANCER DIAGNOSIS

Horacio Emidio de Lucca Junior[1,2] and Arnaldo Rodrigues Santos Jr[2]

[1]Centro Educacional da Fundação Salvador Arena, CEFSA,
São Bernardo do Campo, SP, Brazil
[2]Centro de Ciências Naturais e Humanas (CCNH),
Universidade Federal do ABC, São Bernardo do Campo, SP, Brazil

ABSTRACT

Diseases that are characterized by the disordered growth of cells that, in many cases, have the property of invading tissues and organs are commonly called cancer. Such cells divide quickly and the invasion can be very aggressive and uncontrolled, resulting in formation of malignant tumors Mammographic images from libraries of the American digital database DDSM were used in this research for digital improvement and characteristic analysis using the OpenVino computer program This work has as main objective to analyze mammography images of breast nodules and to propose a method of classification by shape and texture using computer programs that can maximize the accuracy in the correct diagnosis regarding the malignancy or not of a tumor. It is a tool that it can be useful as a contribution in the interpretation of the results to mastologists who identify such nodules through the analyzed radiological images.

KEYWORDS

Diagnostic imaging, Image processing, Computer-Aided Detection, Computer-Aided Diagnosis.

1. INTRODUCTION

The aid to early diagnosis of cancer has been an incessant search by researchers and companies that study analysis of computational images. They are dedicated to developing systems for this and perhaps making millions of lives less traumatic (EADIE et al., 2012) [1]. This research aims to contribute to a better understanding of mammographic images by evaluating the Intel *OpenVino* program, verifying its effective assistance in interpreting the results obtained. Studies indicate that, although breast cancer can affect people of all ages, the main risk factor is age because the rate of increase increases rapidly for patients up to 50 years old. After that age, the increase occurs more slowly (MARX, 2003) [2], but other risk factors such as those related to the woman's reproductive life, family history of breast cancer, in addition to the high density of breast tissue, are considered. Another situation that has also been found to be a risk factor is exposure to ionizing radiation, even at low doses, especially during puberty (INCA, 2019) [3].

On February 22, 2018, an interview by Lavínio Nilton Camarim, then president of the Regional Council of Medicine of São Paulo (Cremesp), was published in Exame Magazine (Editora Abril), reporting that in a survey conducted by them, it was found that 88% of newly graduated doctors

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.77-84, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111206

did not know how to interpret mammography results. Diagnostic errors can also lead to false positives, where patients undergo unnecessary treatments. As the contour of the masses in mammographic examinations are not well defined as to the limits of the images, the use of techniques that are not capable of making precise segmentations can be effective. As seen in the works of Hussain et al (2014) [4], Cheikhrouhou, Djemal and Maaref (2011) [5] where the variations of the derivative signals at different points of interest in the contour of the masses were evaluated or in Rocha et al. (2016) [6], which used levels of diversity and patterns of LBP (local binary patterns) and gray level co-occurrence matrices (GLCM) for the extraction of texture characteristics.

Several computational techniques have been used in order to develop tools that assist in the interpretation of mammographic exams. These include identifying structures compatible with tumors, aiming at improving the rate of early detection of breast cancer (GIGER, 2000) [7]. Although we can see that this theme has been going on for some time, systems that help in the detection of threats, CAD (Computer-Aided Detection), and those that help in the diagnosis of diseases, CADx (Computer-Aided Diagnosis) systems, are already present in several diagnostic imaging centers. There occur especially in developed countries, such as the USA and European countries (TAYLOR et al., 2004; FENTON et al., 2007) [8] [9]. Such techniques are improved over time in order to have the greatest possible effectivity, therefore the objective of this work is to study a technique different from those specified here to analyze an accuracy of this method.

2. MATERIALS AND METHODS

In the area of computing, there are many challenges for collaborating with medical diagnostics using images, since the acquisition of these, going through the pre-processing and segmentation phases to, finally, classify them. The use of auxiliary programs for pattern recognition has increased exponentially in recent years, these recognitions are made better by machines than by humans (NOBESHI, 2016) [10], for professor researcher from USP (University of São Paulo), Dr. Alexandre Chiavegatto Filho, *"It was believed that the greatest transformations in medicine would occur with the use of robots in corridors or surgical centers, the great advance, however, are the systems that recognize patterns in illnesses and offer doctors elements that help them in making decision-making."*

After defining the language of use, in the Python case, the next step was the choice according to the program that helps in the interpretation of the images studied. Some computer visions are important for understanding the functioning of the program that contributes to the classification of images. Convolutional Neural Networks (CNN) are used mainly for image classification, while convolutional filters are used to extract characteristics from images. These are applied in several layers, and at the end of the training, the model learns to distinguish the most important characteristics. There is a need for a large volume of images for more effective classification. Therefore, it is important to transfer learning, that is, use the network that has already been trained and put a new layer on it. In this case, it would be training in mastering mammographic images. The program chosen was OpenCV, an image processing library developed by Intel. The choice of this program was mainly due to the fact that this library is available on Mac, Windows and Linux, works in C, C ++ and Python, and would be a free open source and easy to use and install.

2.1. Image Bank

For the analysis of mammographic images to be carried out effectivity, it was necessary to use an image bank that could corroborate the objectives of the work. Using a standard test database is

important for researchers to compare results directly. The most common databases for the analysis of mammographic images are the database of the Mammographic Image Analysis Society (MIAS) and the Digital Database for Screening Mammography (DDSM). DDSM's main objective is to provide access to images to facilitate research in the development of computational algorithms that can assist in their screening, as well as assist in the diagnosis and development of didactic or even training material. Approximately 2500 studies are included in this database. After analyzing and studying the image banks available for the research, the use of the DDMS database was determined, mainly due to the quality, diversity of incidence and quantity of images. This project was approved by the UFABC Ethics Committee (Process 08/2020).

2.2. Using the program

The images used were in jpg format. First, the <cv.blur> algorithm was designed to simplify the texture and thus give greater projection to the research object, in this case, the breast nodules.

```
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
img = cv.imread('Imagem1.jpg')
blur = cv.blur(img,(10,10))
plt.subplot(121),plt.imshow(img),plt.title('Original')
plt.xticks([]), plt.yticks([])
plt.subplot(122),plt.imshow(blur),plt.title('Blurred')
plt.xticks([]), plt.yticks([])
plt.show()
cv.imwrite('imagem1b.jpg',blur)
```

Figure 1 – commands for *image smoothing*

The following figure shows the original image (Imagem1.jpg) and cv.imwrite (imagem1b.jpg) after highlighting its texture

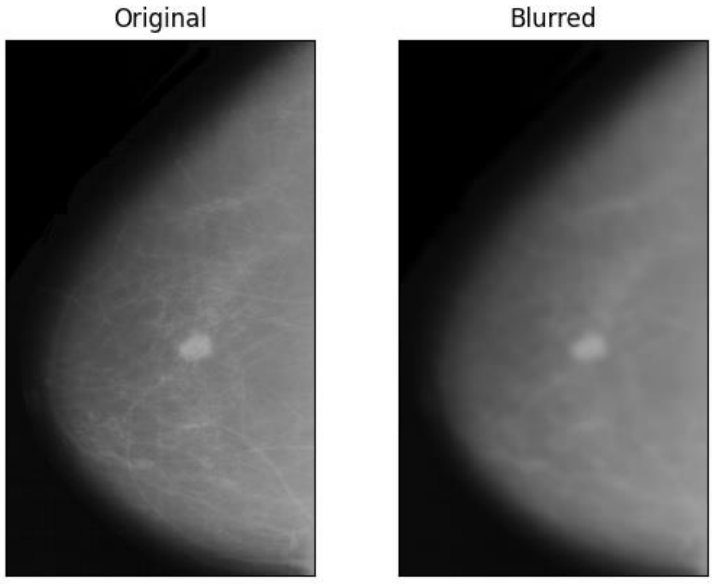


Figure 2 – blurred image obtained from the original image

This process is also called *image smoothing,* it removes high-frequency content normally occurs at the edge contour.
From the *image smoothing* or *blurred image*, the command that determined the segmentations on the edge of the identified tumors was used, observing that the input image is the one obtained by the previous process.

```
import numpy as np
import cv2 as cv
from matplotlib import pyplot as plt
img = cv.imread('imagem1b.jpg',0)
edges = cv.Canny(img,1,3)
plt.subplot(121),plt.imshow(img,cmap = 'gray')
plt.title('Original Image'), plt.xticks([]), plt.yticks([])
plt.subplot(122),plt.imshow(edges,cmap = 'gray')
plt.title('Edge Image'), plt.xticks([]), plt.yticks([])
plt.show()
cv.imwrite('imagem1contorno.jpg',edges)
```

Figure 3 – commands for segmentations on the edge

However, the edges = cv.Canny (img, x, y) parameters must be adjusted, because when x = 1 and y = 3, for example, many contours can be obtained, as shown in figure 4.

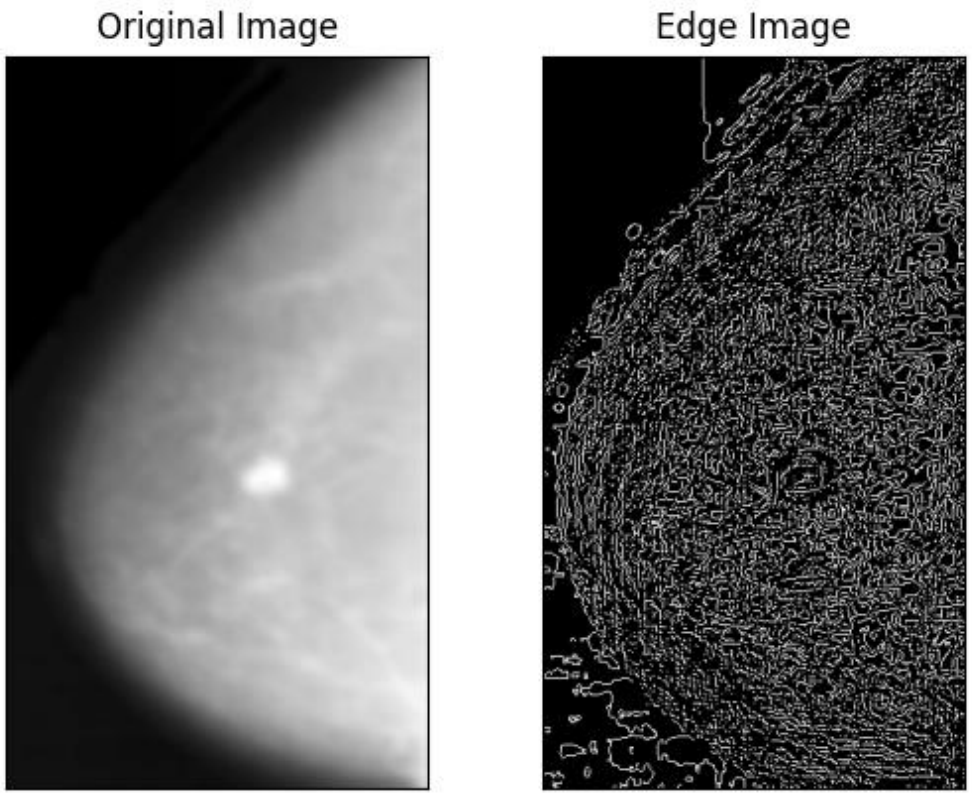


Figure 4 – segmentations on the edge image obtained from the original image (blurred image)

It is the appropriate adjustments of the x and y variants that will determine the proper identification of the nodules. The parameters x = 10 and y = 30 were then used.

```
import numpy as np
import cv2 as cv
from matplotlib import pyplot as plt
img = cv.imread('imagem1b.jpg',0)
edges = cv.Canny(img,10,30)
plt.subplot(121),plt.imshow(img,cmap = 'gray')
plt.title('Original Image'), plt.xticks([]), plt.yticks([])
plt.subplot(122),plt.imshow(edges,cmap = 'gray')
plt.title('Edge Image'), plt.xticks([]), plt.yticks([])
plt.show()
cv.imwrite('imagem1contorno.jpg',edges)
```

Figure 5 – commands for segmentations on the edge with x = 10 and y = 30

Obtaining the appropriate image, as shown in the figure 6, where it was possible to identify the nodule.

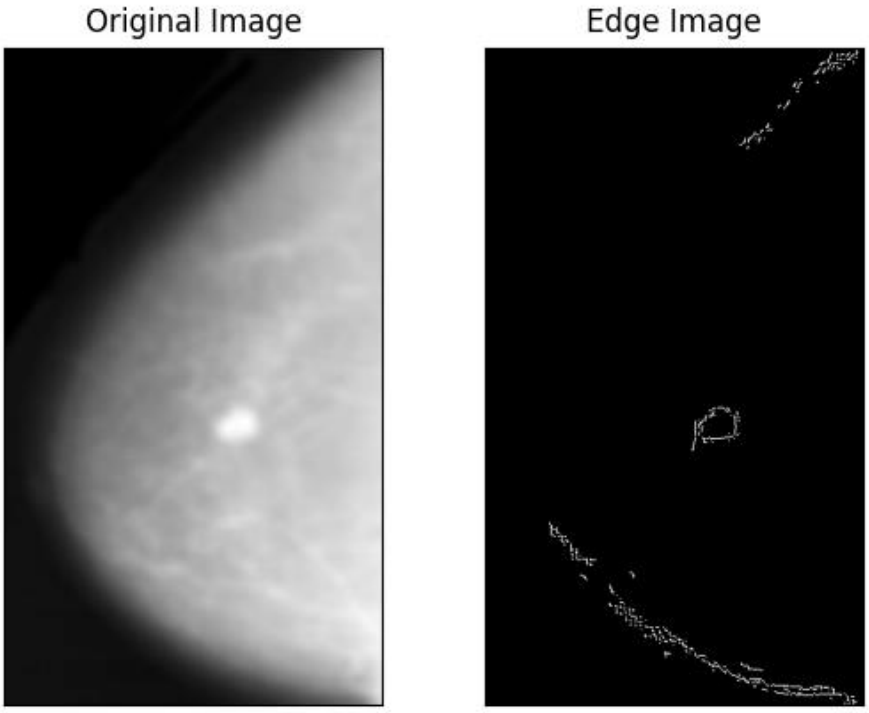


Figure 6 – segmentations on the edge image obtained from the original image (blurred image) with x = 10 and y = 30

After making several variations for x and y, it was concluded that the best results were obtained when $9 < x < 12$ and $28 < y < 31$.

2.3. Results

The mammographic images used were divided by incidence and side of the breast. The parameters x = 10 and y = 30 were then used for the correct identification of the nodule. For the craniocaudal incidence (CC) on the right side, 237 images were used, obtaining an 87.8% effective in the identification and classification of malignant nodules, since they were observed in 208 of the analyzed images.

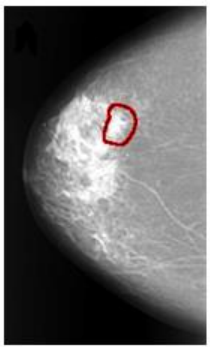


Figure 7 - nodule identified in the skull - caudal view on the right side – (DDMS)

For the left craniocaudal incidence (CC), 243 images were used with 86.8% effective (211 images) in the identification and classification of malignant nodules. From orthogonal views, there was also a classification for mediolateral-oblique view on the right side, with 186 images and 84.4% effective, observed in 157 images.

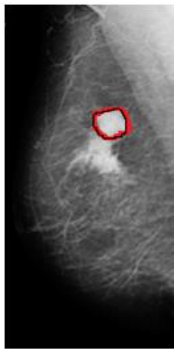


Figure 8 - nodule identified in right lateral mediolateral-oblique image (DDMS)

Finally, 213 mediolateral-oblique images on the left side were also used with 86.9% effective in the identification and classification of malignant nodules, that is, in 185 images.

The same images were analyzed using the parameters x = 11 and y = 29, but the correct results obtained percentages below with these parameters. For the craniocaudal incidence (CC) on the right side, 237 images were used, obtaining an 85.7% effective in the identification and classification of malignant nodules, since they were observed in 203 of the analyzed images. For the left craniocaudal incidence (CC), 243 images were used with 84.4% effective (205 images) in the identification and classification of malignant nodules. From orthogonal views, there was also a classification for mediolateral-oblique view on the right side, with 186 images and 83.3% effective, observed in 155 images, and 213 mediolateral-oblique images on the left side were also used with 85.4% effective in the identification and classification of malignant nodules, that is, in 182 images.

2.4. Discussion

Breast cancer is one of the most aggressive tumors and brings the highest incidence of death among women. Screening by means of mammograms is the main means of early detection for the diagnosis of malignant neoplasms of the breast, one of the main causes of death in different countries (XAVIER et al., 2016; CHOI et al., 2018; BOUJEMAA et al., 2019) [11] [12] [13].

In relation to the radiological technique used, the contrast between the tissue to be identified and the background is a major factor for the perception of breast lesions still in the early stages. The ratio, low electrical voltage or difference in electrical potential (kilovoltage - Kv) and high intensity of electrical current (milliamperage - mA) define the high contrast necessary to obtain the image of the breast. With the program used, it was possible to analyze several characteristics of the images. Additionally, some factors impacted the correct interpretation, such as heterogeneously dense breast tissue, which can hinder the detection of small nodules.

However, it was possible to observe isodense, oval, partially obscured, bilateral nodules, smaller than 1.0 cm, as well as identifications of suspicious microcalcifications, even thin and pleomorphic, with evolutionary forms. With greater ease, the program detected high density spiculated nodules, with diameters greater than 1.0 cm, characteristic of malignant neoplasia. The results obtained were satisfactory in relation to the effectiveness of the method. However, it is worth mentioning that the number of images analyzed was low for a better performance of the program so that they were properly classified in convolutional neural networks. Another factor that implies a more promising result is the fact that the images worked in the DDMS database are images that are more than 15 years old. Therefore, these images are not as sharp as the most recent ones coming from more modern mammographs.

3. CONCLUSIONS

With this work it was possible to conclude that the image training method used with the *OpenVino* program obtained promising results. An average of 86.6% effective was achieved in detecting malignant nodules from mammographic images. It is important to highlight that, in Brazil, a considerable part of the population does not have access to adequate medicine for the treatment, nor for the early diagnosis of the disease. Therefore, using a method that helps the correct interpretation of mammographic exams by doctors and radiologists, contributes more accurately to the correct diagnosis of breast cancer. The effectivity values and the success rate of the tumors are considered relatively good, however the use of more modern images and, consequently, with higher resolution, would give greater precision in the correct diagnosis.

REFERENCES

[1] EADIE, L. H.; TAYLOR, P.; GIBSON, A.P. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. European Journal of Radiology, v. 81, n.1), e70-e76, 2012.

[2] MARX, A. G.; FIGUEIRA, P. V. G. Fisioterapia no câncer de mama. Barueri, SP: Atlas, 2003.

[3] https://www.inca.gov.br/

[4] HUSSAIN, M.; KHAN, S.; MUHAMMAD, G.; AHMED, I.; BEBIS, G. Effective extraction of gabor features for false positive reduction and mass classification in mammography. Appl. Math, v. 8, n. 1L, p. 397-412, 2014.

[5] CHEIKHROUHOU, I.; DJEMAL, K.; MAAREF, H. Protuberance selection descriptor for breast cancer diagnosis. In: IEEE. 3rd European Workshop on Visual Information Processing (EUVIP) 2011. Paris, France, 2011. P. 280-285.

[6] ROCHA, S. V. da; JUNIOR, G. B.; SILVA, A. C.; PAIVA, A. C. de; GATTASS, M. Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution. Expert Systems with Applications, Elsevier, v. 66, p. 7-19, 2016.

[7] GIGER, M. L. Computer-aided diagnosis of breast lesions in medical images. Computing in Science & Engineering, AIP Publishing, v. 2, n. 5, p. 39-45, 2000.

[8] TAYLOR, P.; CHAMPNESS, J.; GIVEN-WILSON, R.; POTTS, H.; JOHNSTON, K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. The British journal of radiology, v. 77, n. 913, p. 21, 2004.

[9] FENTON, J. J.; TAPLIN, S. H.; CARNEY, P. A.; ABRAHAM, L.; SICKLES, E. A.; D'ORSI, C.; BERNS, E. A.; CUTTER, G.; HENDRICK, R. E.; BARLOW, W. E. et al. Influence of computer-aided detection on performance of screening mammography- New England Journal of Medicine, Mass Medical Soc, v. 356, n. 14, p. 1399-1409, 2007.

[10] NOBESCHI, A. Saúde: como a inteligência artificial pode ajudar nos diagnósticos. Available in: https://epoca.globo.com/saude/noticia/2016/12/saude-como-inteligencia-artificial-pode-ajudar-nos-diagnosticos.html. Access in 12/22/2019

[11] XAVIER, D.R.; OLIVEIRA, R.A.D.; MATOS, V.P.; VIACAVA, F.; CARVALHO, C.C. Cobertura de mamografias, alocação e uso de equipamentos nas Regiões de Saúde. Saúde Debate, v.40, n.110, 2016.

[12] CHOI, K.S.; YOON, M.; SONG, S.H.; SUH, M.; PARK, B.; JUNG, K.W.; JUN, J.K. Effect of mammography screening on stage at breast cancer diagnosis: results from the Korea National Cancer Screening Program. Scientific Reports, v.8, 8882, 2018.

[13] BOUJEMAA, S.; BOSMANS, H.; BENTAYEB, F. Mammography Dose Survey Using International Quality Standards. Journal of Medical Imaging and Radiation Sciences, v.50, n.4, p.529-535, 2019.

AUTHORS

Horacio Emidio de Lucca Junior – graduated in Mathematics (2001), graduated in Pedagogy (2015), postgraduate – 360h in Mathematics Education (2004) and Master in Applied Mathematics from Federal University of ABC (2016). Has experience in Mathematics, with emphasis on Analytical Geometry, Calculus, Financial Mathematics, Statistics and Probability. PhD student in Biotechnology at the Federal University of ABC (conclusion 2022)

Arnaldo Rodrigues Santos Junior - graduated in Biological Sciences (1993), Master's degree in Cell Biology (1996) and a PhD in Cell and Structural Biology (2001). Has experience in the area of Morphology, with emphasis in Cytology and Cell Biology. It works on the following topics: cell culture, cell growth and differentiation, biomaterials, morphological analysis, bioresorbable polymers and Biotechnology.

DETECTION OF OIL TANK FROM HIGH RESOLUTION REMOTE SENSING IMAGES USING MORPHOLOGICAL AND STATISTICAL TOOLS

D. Chaudhuri[1] and I. Sharif[2]

[1]DRDO Integration Centre, Panagarh Base, Muraripur,
Burdwan – 731219, W.B., India
[2]Defence Research & Development Establishment (DRDE), Jhansi Road,
Gwalior-474002, Madhya Pradesh, India

ABSTRACT

Oil tank is an important target and automatic detection of the target is an open research issue in satellite based high resolution imagery. This could be used for disaster screening, oil outflow, etc. A new methodology has been proposed for consistent and precise automatic oil tank detection from such panchromatic images. The proposed methodology uses both spatial and spectral properties domain knowledge regarding the character of targets in the sight. Multiple steps are required for detection of the target in the methodology – 1) *enhancement technique using directional morphology*, 2) *multi-seed based clustering procedure using internal gray variance* (*IGV*), 3) *binarization and thinning operations*, 4) *circular shape detection by Hough transform*, 5) *MST based special relational grouping operation and* 6) *supervised minimum distance classifier for oil tank detection. IKONOS and Quickbird satellite images are used for testing the proposed algorithm. The outcomes show that the projected methodology in this paper is both precise and competent.*

KEYWORDS

Recognition, remote sensing, resolution, enhancement, supervised procedure, clustering, minimal spanning tree (*MST*).

1. INTRODUCTION

Extractions of features from satellite based images are significant assignment in various applications. Automatic detection of geographic objects such as bridges [1], buildings [2]-[5], ocean disturbance features [6], roads [7] etc from satellite images is useful in many essential purposes including the construction and preservation of correct physical databases, evaluation of the degree of destruction after natural calamity such as floods or earthquakes and military operations.

Few automatic oil tank detection algorithms are accessible in the literatures [8], [16]. Among them, template matching [9] and Hough transform [10]-[12] are common methodologies. There are some drawbacks in the template matching technique – a) time consuming and b) choice of template. Shape is an important feature for target detection and oil tank is mostly a circular shape feature in the image, which may detect by Hough transform [10], [12]. Due to noise or poor

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.85-97, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111207

illumination Hough transform failed to detect oil tanks very well unless good quality enhancement technique is imposed for isolation the foreground and background.

Multiple steps technique – image fusion, Canny edge detector Hough transform and fast matching method is proposed by Zhang *et al*. [13]. Wang *et al*. [14] suggested a very simple method using SAR, which is used for bright spot detection and optical image, which is used for detection of shape feature of all those bright spot. This method [4] is depended on both SAR and visible band images in the same region and this is the major drawback of this method because extra step for registration is required, which is critical for such kind of images. Another technique based on image fusion and morphological operations is suggested by Qiang *et al*. [15]. Kushwaha *et al*. [16] proposed a method to extract bright oil tanks by morphological operation. But the method is unable to detect dark oil tanks. Xuan and Yunqing [25] proposed an oil tank detection technique by three steps approach – i) advance visual saliency model are used for separation of oil tanks from the heterogeneous background, ii) detected circular shadow regions and iii) graph search method and prior knowledge are applied for removal of false oil tanks. Recently Zalpour *et al*. [26] proposed a multi-steps approach for oil tank detection. First they have extracted ROI by using R-CNN and then circles are detected from ROI. Next, features are extracted by using CNN and HOG feature extractors. Lastly, for classification of oil tanks they have used SVM. The method is computationally expensive.

The proposed method is to detect oil tanks from high resolution PAN images using various steps. The basic steps are: 1) directional morphological enhancement module to enhance the required objects, 2) detection and clustering the internal gray variance (IGV), 3) binarization and thinning operation, 4) circular object detection by Hough transform, 5) MST based spatial relationship grouping operation and 6) supervised classification based on statistical and texture features. The algorithm is tested on a variety of images and the results are adequate. Also we have compared our method with Xuan and Yunqing [25] method. It has been observed that two oil tanks with low intensity values of an image are not able to detect by Xuan and Yunqing [25] method. Same oil tanks are detected by the proposed algorithm because the advance directional enhancement technique is able to highlight beautifully both the oil tanks. As a result we are able to isolate these targets from the background by the proposed multi-seed based clustering technique. The remainder of the paper is organized as follows. The proposed oil tank detection methodology is presented in Section 2. Experimental results are discussed in Section 3. We conclude with a summary in Section 4.

2. OIL TANK DETECTION METHODOLOGY

The detection of oil tank from high-resolution PAN images is the centre of focus. The research is very useful for defence aspects as well as civilian aspect. Normally segmentation is one of the major steps in the target detection problem. In the proposed technique we are not segmented the target from the background. We are more concentrated in the enhancement technique by which the targets are more prominent than the background. Then we have transformed the enhanced image into feature domain and then we developed a clustering technique of the feature image for isolation possible oil tanks. Finally, three important step properties are used – shape property, neighbouring relational property and supervised classification property. The overall methodology of the proposed algorithm is shown in Figure 1. Each step is described in the next subsequent sections.

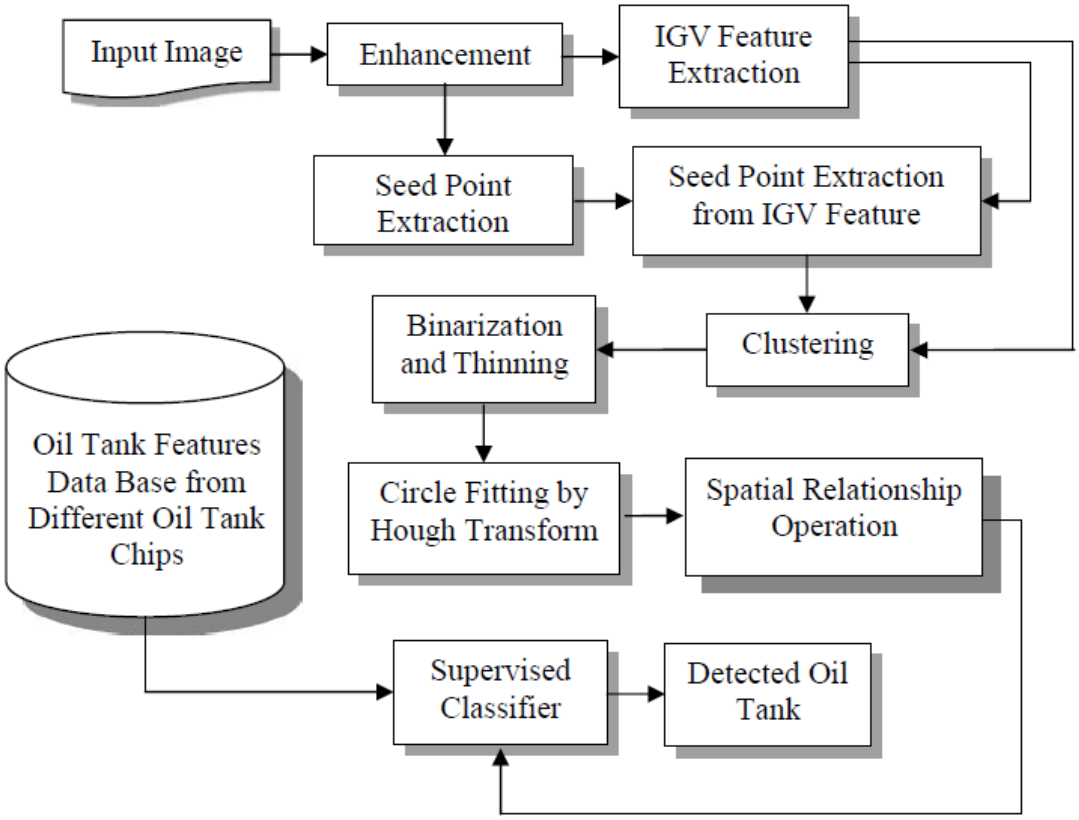


Figure 1. Schematic of the proposed algorithm

2.1. Image Enhancement

Better visualization effect and better contrast between the object and non-object in the image are the main aspect of enhancement technique. The purpose of enhancement technique is that object of interest is more homogeneous than non-interested objects. A good enhancement technique will help for isolation of interested object from the background. Morphological filter [17, 18] is an effective tool for such type of operation. A novel directional morphological filtering technique has been proposed by Chaudhuri *et al.* [7] for enhancement of the object of interest. The same enhancement technique [7] has been applied in the present paper. First select the template of size 5×5 of the object, which is called the structuring element. We have considered four directions horizontal (0^0), right diagonal (45^0), vertical (90^0) and left diagonal (135^0) within a 5×5 window. The proposed enhancement is the directional morphological technique. The choice of direction can be determined by computing standard deviation among the four directions within the 5×5 window of the original image is minimum. Alternate sequence of opening and closing operations are applied to the original image within the 5×5 window by using the structuring element along the minimum directional standard deviation. This operation is followed for the whole image and obtained the final enhanced image.

2.2. IGV Feature Extraction

The region of interest is more homogeneous than the border part of the region by the previous image enhancement technique. If the region or object in the image is homogeneous then the internal gray variance (IGV) [2] will be very small and consequently in the border region, IGV will be higher than interior body of the object.

Here, a moving window $W \times W$ (W = 5) is considered and compute the mean gray value of the image at the present window position. The candidate pixel is the middle pixel of the window and the value is replaced by the square sum difference from each pixel from mean value of the image at present window position. It has been observed that the deviation at the border region of the manmade object (like, oil tank) is high than the natural object.

2.3. Seed Point Extraction Technique

The isolation of manmade and natural objects from a scene of the panchromatic image is very difficult task due to unknown classes of materials, overlie of the brightness values of the classes and multi-modal characteristics of a particular class. As a result, it is very hard job to segment the unprocessed image, particularly synthetic and non-synthetic (natural) classes. To defeat this we have changed the enhanced image into IGV attribute space and then a seed based clustering technique has been applied to separate the artificial object from the rest image.

The basic concept of clustering based on seed point is selection of initial seed point and grow cluster surround the seed. Most beautiful nature of seed based cluster is that it can separate 100% correct if the data classes are circular and well separated. It is unable to cluster correctly if the data is elongated or complex. Chaudhuri *et al*. [22], [23] proposed multi-seed concept to overcome this situation.

Present seed point detection technique is based on our previous work [2]. Here, seed points are detected from the domain information of the enhanced image and IGV feature values. First, seed points are extracted using a multi-seed technique of enhanced image, which is described in Algorithm A. Lastly, the final seed points of IGV feature values, *variance seeds* (*VS*) [2] are extracted with the help of the seed points of enhanced image as Algorithm B. IGV feature space is clustered by using *VS*, which is described in Section D. The interested reader can go through our previous work [2] and here both the algorithms are described for better understanding.

Algorithm A:

Step 1: Input image is the enhanced image (current region, *I*). Find $gr_{\min}$ and $gr_{\max}$ are the minimum and maximum gray values. Let gr_i and corresponding hr_i, i=1,2,…,V are the gray value of and frequency, respectively.

Step 2: Compute *mode* of region *I* i.e. $m = \max_{j=gr_{\min}}^{gr_{\max}} \{hr_j\}$.

Step 3: Find standard deviation (*Std*) w. r. t. mode (m) for region, *I*. That is,

$$Std = \left[\frac{1}{\sum_{i=gr_{\min}}^{gr_{\max}} hr_i} \sum_{i=gr_{\min}}^{gr_{\max}} (m - gr_i)^2 hr_i \right]^{1/2}$$

Step 4: Select threshold T_1 (*homogeneity factor*) and if $Std > T_1$ then follow Step 5. Otherwise follow Step 8.

Step 5: Select the parameter ν, *Gaussian multiplier* and accumulate the set of pixels with gray values gr_i satisfy the constrains, $m - \nu \times Std \le gr_i \le m + \nu \times Std$.

Step 6: Eliminate all $gr_i \in [m - \nu \times Std, m + \nu \times Std]$ from region *I*. Residual pixel gray values are separated into two sets – i) $gr_i \in [gr_{\min}, m - \nu \times Std)$ and ii) $gr_i \in (m + \nu \times Std, gr_{\max}]$.

Step 7: If both or any one of the set in Step 6 are non-homogeneous or bigger set then reiterate Steps 1-6.

Step 8: Stop.

Suppose the extracted *k* seed points by Algorithm A are $m_1, m_2, ..., m_k$; which are mode of corresponding *k* homogeneous clusters, C_j, $j = 1,2,...k$. Next these seed points and corresponding clusters $(m_j, \{C_j\})$ will be utilized in the IGV feature space (Algorithm B) for extraction *k* seed points in the feature domain.

Algorithm B:

- Find the set $PX_j = \{(IV_{xj}, IV_{yj})\}$ from IGV feature space for all pair $(m_j, \{C_j\})$ of the image space, for $j = 1,2,\ldots,k$, $x = 1,2,...,R$ and $y = 1,2,...,S$ where R and S are the rows and columns of the image.
- Extract *IGV* values for corresponding (IV_{xj}, IV_{yj}) i.e. $IGV(IV_{xj}, IV_{yj})$, $j = 1,2,\ldots,k$.
- Compute $GV[j] = \sum_{\forall (IV_{xj}, IV_{yj}) \in PX_j} IGV(IV_{xj}, IV_{yj}),\ j = 1,2,...,k$

Compute seed in the IGV feature space, $IVS[j] = \frac{GV[j]}{\#PX_j},\ j = 1,2,...,k$ where "$\#PX_j$" is the total number of points of the group PX_j.

2.4. Clustering Technique

The proposed clustering technique is based on multi-seed clustering technique [22], [23]. It has been observed that many small clustered regions are formed after applying the proposed clustering technique to enhanced image and identification oil tank is a hard job from such clustered image. This problem can be handled by nearest neighbor clustering technique in IGV feature space using $VS[j], j = 1,2,...,k$. The manmade structures are isolated in a single group by applying this clustering technique. So, the automatic threshold detection is very easy from such clustered image for formation of binary image.

2.5. Threshold and Thinning Techniques

The previous IGV feature space cluster image is a gray level image with limited gray variance and it is very easy to convert a binary image for isolation of manmade and natural objects by simple threshold technique, provided appropriate selection of threshold value. There are various automatic threshold value detection techniques in the literature. In this paper, bimodality detection approach [23] is used for detection of automatic threshold value.

The data is said to be bimodal if the data can be divided into two sub-data. Suppose *Q* be the data population of the IGV cluster image. *Q* is divided into two parts, say $Q_{L(u)}, Q_{H(u)}$ in such a way so that i) $Q_{L(u)}$ includes all the data with group value ≤ *u*, and $Q_{H(u)}$ includes all the data with group value > *u*. and ii) the variances of $Q_{L(u)}$ and $Q_{H(u)}$ are small in respect the variance of *Q*.

Suppose N and σ^2 are the total data frequency and variance of Q, $N_{L(u)}$ and $\sigma^2_{L(u)}$ be the data frequency and variance of $Q_{L(u)}$ and lastly, $N_{H(u)}$ and $\sigma^2_{H(u)}$ be the data frequency and variance of $Q_{H(u)}$. Now for estimation the threshold value, say $u = T$ (*bimodality parameter*), we will construct the objective function, which will be minimum among all the gray values of IGV cluster image as

$$OB(u) = \frac{N_{L(u)}\sigma^2_{L(u)} + N_{H(u)}\sigma^2_{H(u)}}{N\sigma^2}$$

So T will be the threshold value for thresholding IGV cluster image, (IGV_{Cl}) to binary image as

$$IGV_{Binary}(x,y) = \begin{cases} 0 \text{ if } IGV_{Cl}(x,y) < T \\ 1 \text{ if } IGV_{Cl}(x,y) \geq T \end{cases}$$

Oil tank is the circular shape object and the edge information is very important for finding the shape. Though IGV_{Binary} image gives the edges of the manmade object but the edges are thicker and thinning algorithm, which is available in the literature [24] is applied for finding the proper edge.

2.6. Circle fitting

Feature extraction in images is one of the most challenging tasks in computer vision. Practically, objects of interest may appear in different shapes and sizes and a solution of this problem is to find an algorithm for extraction any shape and size within an image. Then the objects can be classified accordingly to parameters needed to describe the shapes and most effective method for this is the Hough Transform [12], [19]. In this paper we have extracted circular shapes objects from the binary image by using Hough transform.

2.7. MST Based Spatial Relationship Grouping Operation

Oil tank is an important defence target and normally they formed in a group i.e. many oil tanks are constructed in an open space connected with roads. Isolate tank is not so much important because that may be water tank or other object. In this paper, we proposed Minimal Spanning Tree (MST) based clustering technique for formation cluster of oil tanks. The idea is very simple. Suppose K number of circular shape objects, $OC_i, i = 1,2,..., K$ are detected by the previous steps. First compute the centroids of all the circular shape objects, which are the nodes of the graph and construct the MST between all the centriods. The edge weights between the two connected nodes are computed by the Euclidean distance between two connecting nodes. These edge weights will make the decision that whether the connected nodes will form a group or not. Suppose T_{Gp} is the distance threshold value, which depends on the resolution of the image and the decision logic is if the edge weight between the two connected objects is greater than T_{Gp} then those objects are formed different clusters; otherwise they are in the same cluster. Experimentally we have seen $T_{Gp} = 12$ gives the good result for IKONOS and QuickBird data.

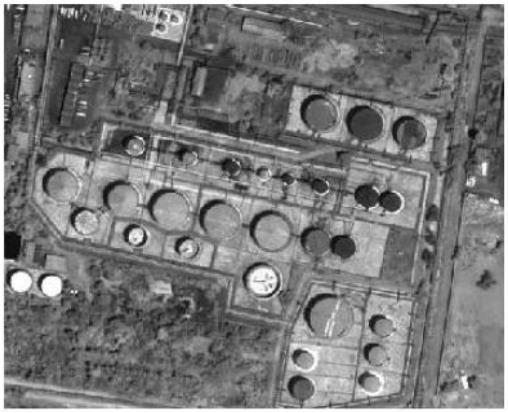
2.8. Supervised classification

The possible oil tanks along with other similar objects are detected from the previous subsections. A supervised classifier has to be develop for detection of confirm oil tanks. Image target chips of oil tanks are stored as training samples and 15 different statistical and texture features [20] are

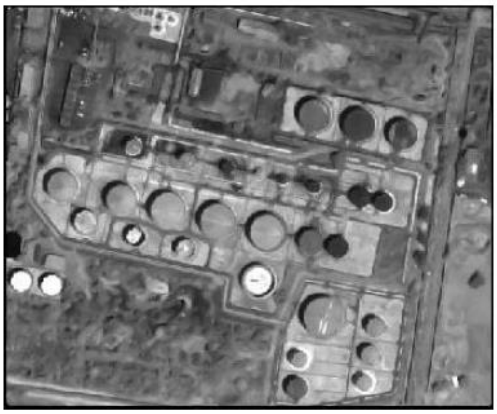
extracted from those training sample oil tank chips. So, each oil tank chip is having a 15 dimensional feature vector space, O_F . Now for the classification, we have extracted the same 15 dimensional features from each possible oil tanks of the output image of the previous step and compute the distances between the 15 dimensional feature vector of the possible oil tank and the training data base O_F . Find the minimum distance among all distances and if the minimum distance is less than the threshold value, T_C then the object is a confirm oil tank. At the same time, we have seen the minimum distance between the similar object feature vector, which is not an oil tank and training data base vector O_F is greater than the threshold value, T_C . In our case, T_C =0.4 is a good choice for IKONOS and QuickBird data.

3. EXPERIMENTAL RESULTS

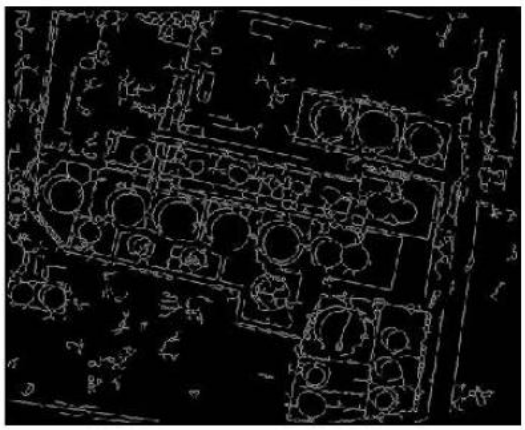
We have tested the various steps of the proposed algorithm in panchromatic images of IKONOS and QucickBird satellites. All the steps of the algorithm are executed and timed on an HP xw6400 workstation (Intel(R) Xeon(R), 5130 at 200 GHz, 2.00 GB RAM, Microsoft windows XP). Total computational cost of oil tank detection by the proposed algorithm is 312 milliseconds.



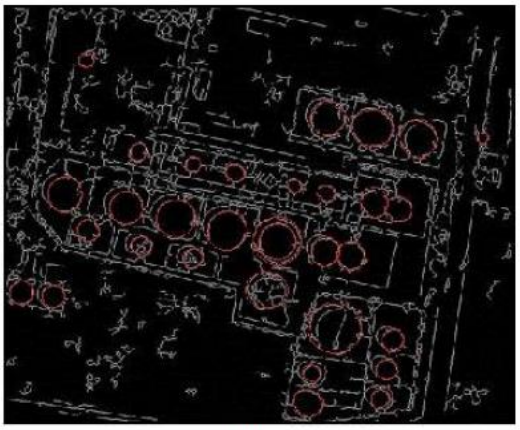
(a)



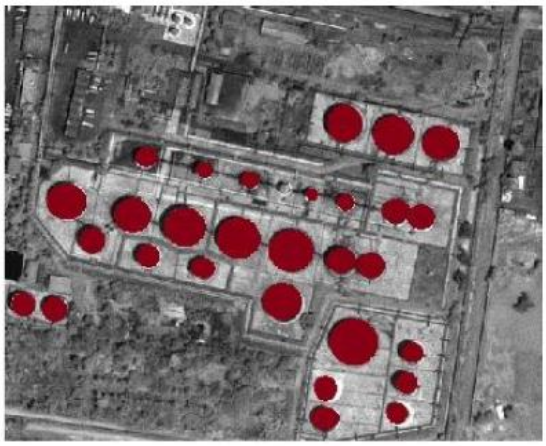
(b)



(c)



(d)

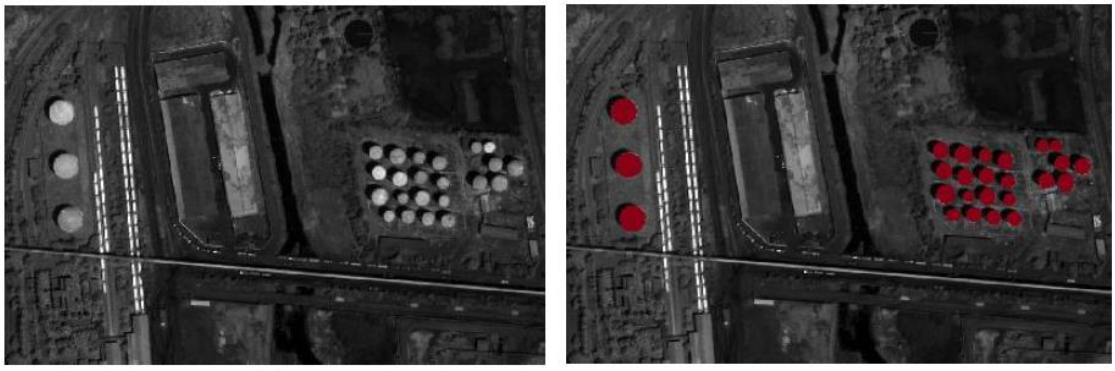


(e)

Figure 2. IKONOS Panchromatic image: (a) original image, (b) enhanced image, (c) binary image, (d) circular shape detected image, and (e) final output

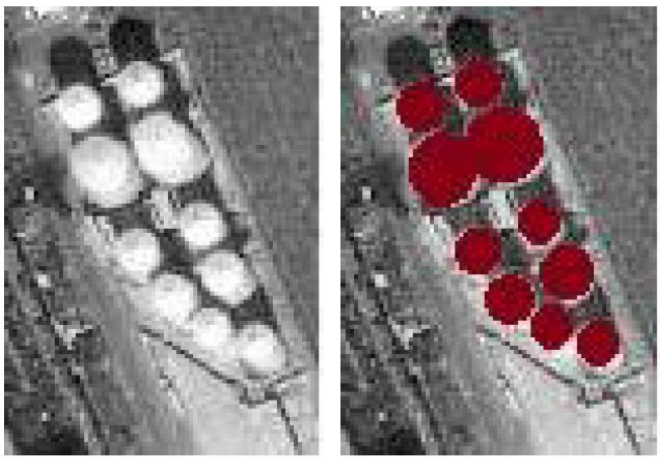
The original IKONOS satellite image of size 493×401 is shown in Figure 2 (a). The proposed enhancement technique has been applied to the original image and Figure 2(b) shows the enhanced image. It is noticed that manmade structures in Figure 2(b) are more uniform and enhanced than the original structures in Figure 2(a). Figure 2(c) shows the segmented image by using the proposed clustering technique of the IGV feature space. The detected circular shape objects by using Hough transform is shown in Figure 2(d). Few flash alarms are appeared in Figure 2(d) and removing those unwanted objects final output is shown in Figure 2(e).

Figure 3(a) and Figure 4(a) show the PAN images from different satellites IKONOS and QuickBird of sizes 910×610 and 78×109, respectively. The detected oil tanks are shown Figure 3(b) and Figure 4(b), respectively by the proposed algorithm.



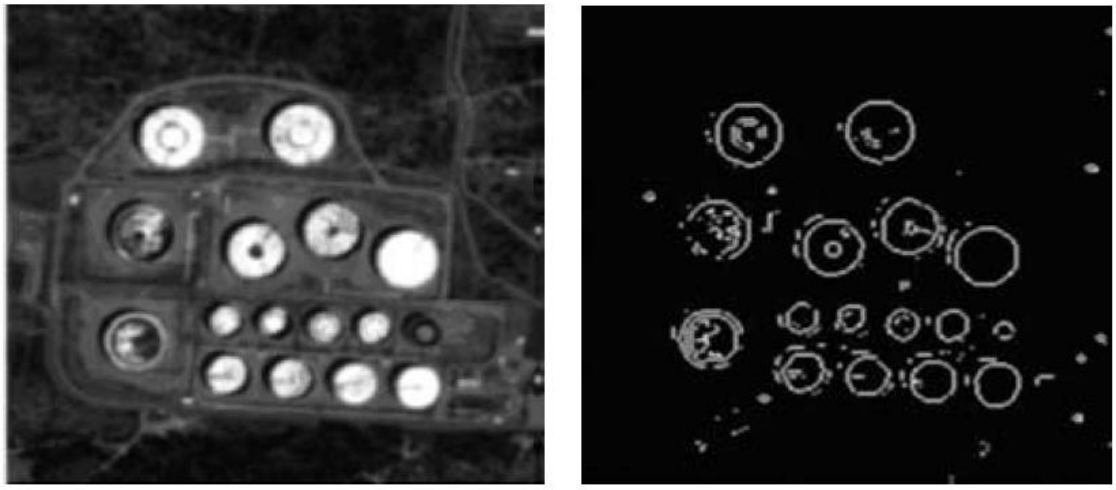
(a) (b)

Figure 3. IKONOS PAN image: a) original image and (b) output image

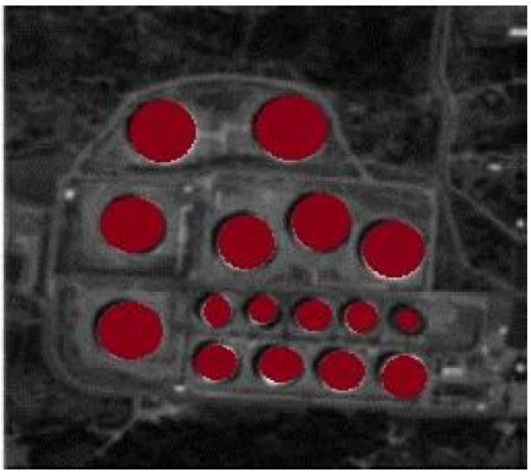


(a) (b)

Figure 4. QuickBird PAN image: (a) original image and (b) output image



(a) (b)



(c)

Figure 5: Original IKONOS image: (a) Original image of size 254×225, (b) detected oil tanks by Xuan and Yunqing [25] and (c) detected oil tanks by the proposed method

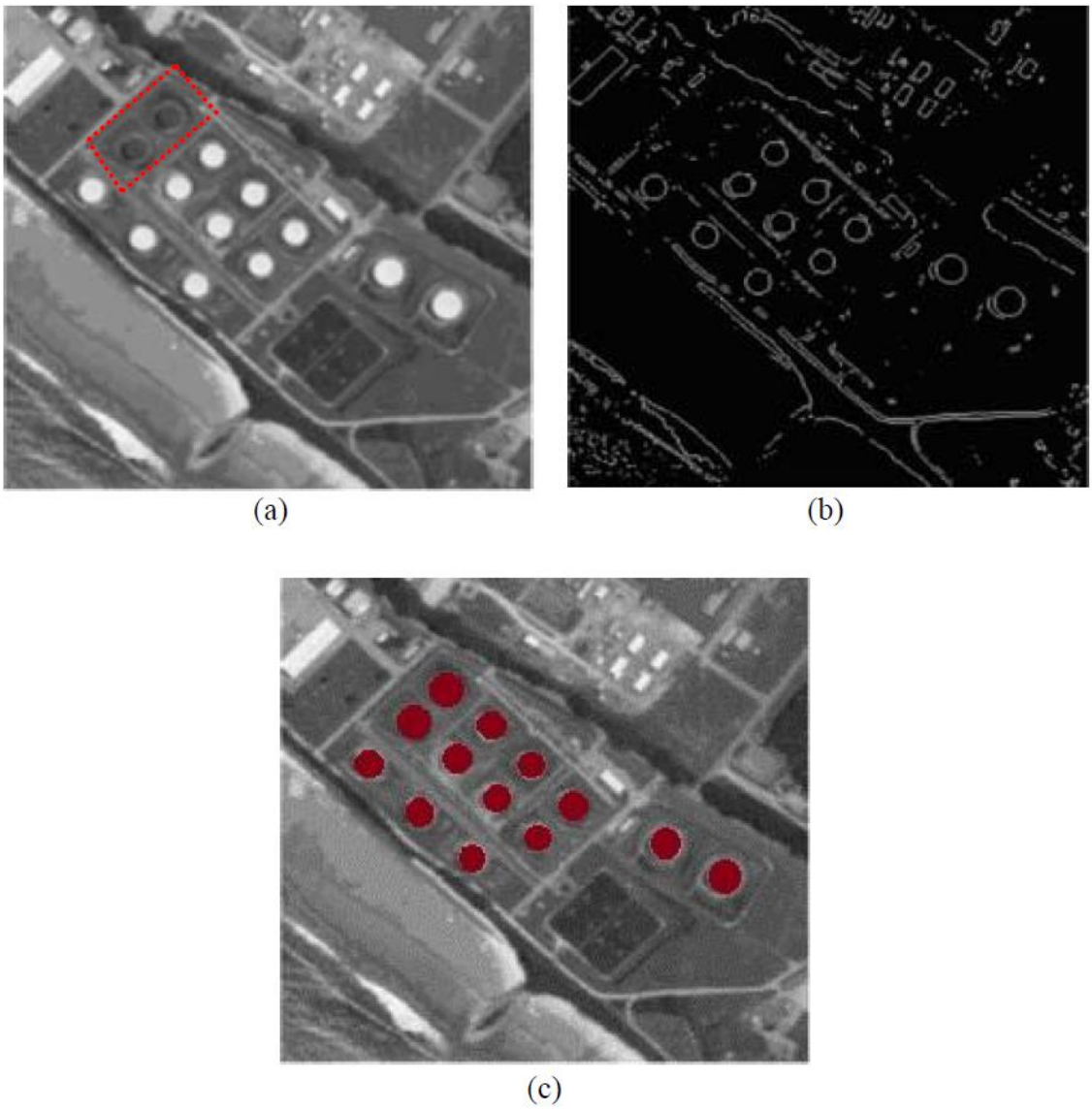


Figure 6: Original IKONOS image: (a) Original image of size 246×226, (b) detected oil tanks by Xuan and Yunqing [25] and (c) detected oil tanks by the proposed method

Figure 5(a) shows an original IKONOS image of size 254×225. The detected oil tanks by Xuan and Yunqing [25] and the proposed method are shown in Figures 5(b) and 5(c), respectively. We have seen that both the methods are detected oil tanks correctly.

Another IKONOS image is shown in Figure 6(a) of size 246×226. The detected oil tanks by Xuan and Yunqing [25] and proposed method are shown in Figures 6(b) and 6(c), respectively. Here we have seen that two black oil tanks inside the dotted red color rectangular shape in the original image Figure 6(a) are not detected by Xuan and Yunqing [25] method. At the same time, both the oil tanks are detected by the proposed technique. It is because of enhancement technique which is highlighted both the oil tanks perfectly and subsequent clustering technique of the proposed method is able to isolate these targets from the background.

4. CONCLUSIONS AND FUTURE SCOPE

Multi-steps algorithm have described in this paper for automatic extraction of oil tank from high resolution PAN satellite images. This problem has many applications for civilian, commercial, and military domains. The significant modules in the projected algorithm are: image enhancement for highlight the target from the background, multi-seed based clustering technique using internal gray variance, binarization and thinning operation, circular object detection using Hough transform, MST based spatial relationship cluster formation and supervised classification using statistical and texture features. A huge amount of training images of the various oil tanks have been generated and 15 dimensional statistical and texture features have extracted for supervised classification module. The proposed algorithm was tested on IKONOS and QuickBird satellites PAN images and satisfactory results have reported by the proposed algorithm. Also we have compared our results with other method and it has observed that the proposed algorithm gives the better results.

Presently machine learning is an advance method for target detection in supervised manner. Different high resolution sensor data are available. Our future work includes the development of algorithms to identify oil tanks from panchromatic high-resolution satellite imagery by using machine learning.

REFERENCES

[1] Chaudhuri, D. & Samal A., (2008) "An automatic bridge detection technique for multi-spectral images," *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 46, No. 9, pp. 2720-2727.

[2] Chaudhuri, D., Kushwaha, N. K., Samal, A. & Agarwal, R. C., (2016) "Automatic building detection from high-resolution satellite images based on morphology and internal gray variance", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9, No. 5, pp. 1767-1779.

[3] Theng, L. B., (2006) "Automatic building extraction from satellite imagery", *J. Engineering Letters*, Vol. 13, No. 3, EL_13_3_5 (Advance online publication), 4 November.

[4] Xiaoying, J. & Davis, C. H., (2005) "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual and spectral information", *Journal of Applied Signal Processing*, Vol. 14, pp. 2196 – 2206.

[5] Khoshelham, K., Nardinocchi, C., Frontoni, E., Mancini, A. & Zingaretti, P., (2010) "Performance evaluation of automated approaches to building detection in multi-source aerial data", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 65, pp. 123 – 133.

[6] Chaudhuri, D., Samal, A., Agrawal, A., Sanjay, Mishra, A., Gohri, V., & Agarwal, R. C., (2012) "A statistical approach for automatic detection of ocean disturbance features from SAR images", *IEEE Journal of selected topics in Applied Earth Observations and Remote Sensing*, Vol. 5. No. 4, pp. 1231 – 1242.

[7] Chaudhuri, D., Kushwaha, N. K., & Samal, A., (2012) "Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation technique", *IEEE Journal of selected topics in Applied Earth Observations and Remote Sensing*, Vol. 5, No. 5, pp. 1538 – 1544.

[8] Zhu, C., Liu, B., Zhou, Y., Yu, Q., Liu, X., & Yu, W., (2012) "Framework design and implementation for oil tank detection in optical satellite imagery", *IEEE International Geoscience and Remote Sensing Symposium*, THP.P11, July, Munich, Germany, pp. 22 – 27.

[9] Gonzalez, R. C. & Woods, R. E., (2008) *Digital Image Processing*, Pearson Education, Delhi, India.

[10] Illingworth, J. & Kittler, J., (1988) "A survey of the Hough transform computing", *Computer Vision, Graphics and Image Processing*, Vol. 44, No. 1, pp. 87 – 116.

[11] Wei, Y., (1998) "Improved dynamic broad sense Hough transform and its application on round detection", *Surveying and Mapping Information and Engineer*, Vol. 4, pp. 23 – 26.

[12] Wu, M., Song, Z., Li, B., Li, F., Li, B., & Shen, C., (2015) "A method to detect circle based on Hough transform", *Int. Conf. on Information, Machinery, Materials and Energy* (ICISMME), Atlantis Press, pp. 2028 – 2031.

[13] Zhang, W., Zhang, H., Wang, C., & Wu, T., (2008) "Automatic oil tank detection algorithm based on remote sensing image fusion", *IEEE International Geoscience and Remote Sensing Symposium* (IGARSS), Vol. 6, pp. 3956 – 3958.
[14] Wang, Y., Tang, M., Tan, T., & Tai, X., (2004) "Detection of circular oil tanks based on the fusion of SAR and optical images", *Third International Conference on Image and Graphics* (ICIG), pp. 524 – 527.
[15] Qiang, Z., Du, X., & Sun, L., (2011) "Remote sensing image fusion for dim target detection", *International Conference on Advanced Machatronic Systems* (ICA Mechs), pp. 379 – 383.
[16] Kushwaha, N. K., Chaudhuri, D. & Singh, M. P., (2013) "Automatic bright oil circular type oil tank detection using remote sensing images", *Defence Science Journal*, Vol. 63, No. 3, pp. 298 – 304.
[17] Haralick, R. M., Sternberg, S. R., & Zhuang, X., (1987) "Image analysis using mathematical morphology", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, pp. 532 – 550.
[18] Serra, J., (1982) *Image Analysis and Mathematical Morphology*, New York: Academic.
[19] Duda, R. & Hart, P. E., (1972) "Use of the Hough transform to detect lines and curves in pictures", *ACM*, Vol. 15, No. 11, pp. 11 – 15.
[20] Haralick, R. M., Shanmugam, K., & Dinstein, I., (1973) "Texture features for image classification", *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 3, pp. 610 – 621.
[21] Cormen, T. H., Leiserson, L. E., Rivest, R. L., & Stelin, C., (2002) *Introduction to algorithm*, Prentice Hall of India Pvt. Ltd., Eastern Economy Edition, New Delhi, pp. 570 – 573.
[22] Chaudhuri, D. & Chaudhuri, B. B., (1997) "A novel multi-seed non-hierarchical data clustering technique," *IEEE Trans. on System, Man and Cybernetics*, Part – B, Vol. 27, No. 5, pp. 871-877.
[23] Chaudhuri, D. & Agrawal, A., (2010) "Split-and-merge procedure for image segmentation using bimodality detection approach", *Defence Science Journal*, Vol. 60, No. 3, pp. 290-301.
[24] Rosenfeld, A. & Kak, A., (1982) *Digital Picture Processing*, New York, Academic.
[25] Xuan, L. & Yunqing, L., (2016) "Oil Tank Detection in Optical Remote Sensing Imagery Based on Quasi-circular Shadow", *Journal of Electronics and Information Technology*, Vol. 38, No. 6, pp. 1489-1495.
[26] Zalpour, M., Akbarizadeh, G. & Alaei-Sheini, N., (2020) "A new approach for oil tank detection using deep learning features with control false alarm rate in high-resolution satellite imagery", *International Journal of Remote Sensing*, Vol. 41, No. 6, pp. 2239-2262

AUTHORS

Dr. Debasis Chaudhuri received the B. Sc. (with Honors) in Mathematics from Visva-Bharati University, Santiniketan, the M. Sc. in Applied Mathematics from Jadavpur University, Kolkata, and Ph. D. degree in image processing and pattern recognition from Indian Statistical Institute, Kolkata. He is currently a senior scientist and DGM at DRDO, Panagarh, India. He was a research associate of Council of Scientific and Industrial Research (CSIR) at Indian Statistical Institute, Kolkata. Dr. Chaudhuri was a project scientist at Indian Statistical Institute in ISI-ADRIN Project (collaborated by ISI and Dept. of Space, Govt. of India). He was a Scientist at Defence Electronics Applications Laboratory, Dehradun, India. He was an associate professor of Defence Institute of Advance Technology (DIAT), Pune. He was a visiting professor at University of Nebraska, USA in the Dept. of Computer Science and Eng. for 2003-2004. He has also visited many other Universities and Institutes within and outside India for delivering invited lectures. He has been a member of the program/organizing committees of many national and international conferences. He has more than thirty years experience in the field of image processing, pattern recognition, computer vision and remote sensing with extensive systems development and implementation experience. He has got lot of experience in advising and guiding others in related fields. He has extensive experience in the field of automatic target detection from satellite imagery. He had successfully completed six national important projects for the services. He had guided many Ph. D/M. Tech/M.S Student/B. Tech students. Two students were awarded the best project award from IEEE at Western Region Software Park, Pune and National Institute of Technology, Nagpur. He has authored or co-authored over 60 papers in international journals and conferences in the area of image processing and pattern recognition. He is reviewer in many international journals. Also he is an Associate Editor of several international journals. He is a senior member of IEEE and fellow of IETE. He has received Technology Award from DRDO Science Forum, Mins. of Defence, Govt. of India in 2011 for excellent contribution in "**Target Detection from Remote Sensing Satellite Data**". He has received "Group Technology Award –

2013" from DEAL, DRDO, Mins. of Defence, Govt. of India for contribution to success the "Image Intelligence Environment" project. He has received Technology Award from DRDO Science Forum, Mins. of Defence, Govt. of India in 2018 for excellent contribution in "**Supervised classification of multispectral images by minimum distance with majority must be granted logic using multi-seed technique"**. His research interests include image processing, pattern recognition, computer vision, remote sensing and target detection from satellite, SAR, Thermal and MMW imageries.

Dr. Imran Sharif received his B. Tech (Computer Science and Engineering) from Institute of Engineering and Technology, Kanpur in 2002 and M. Tech from CDAC, Noida in 2005 and Ph. D. (Computer Science & Engineering) from Uttarakhand Technical University, Dehradun in 2018. Currently he is working as scientist at DRDE, Gwalior. More than 15 years research experience in the field of Hyper spectral image processing, Pattern recognition, Computer vision, Remote sensing and target detection from satellite imagery and Embedded system. He is a life time member of computer society of India and Indian Society of Remote Sensing.

THE ADOPTION OF THE INTERNET OF THINGS FOR SMART AGRICULTURE IN ZIMBABWE

Tsitsi Zengeya[1], Paul Sambo[1] and Nyasha Mabika[2]

[1]Great Zimbabwe University, Department of Mathematics and Computer Science
[2]Great 2Zimbabwe University, Department of Livestock, Wildlife and Fisheries

ABSTRACT

Zimbabwe has faced severe droughts, resulting in low agricultural outputs. This has threatened food and nutrition security in community sections, especially in areas with low annual rainfall. There is a growing need to maximize water usage, monitor the environment and nutrients, and temperatures by the adaptation of smart agriculture. This research explored the use of the Internet of Things (IoT) for smart agriculture in Zimbabwe to improve food production. The mixed methodology was used to gather data through interviews from 50 purposively sampled A2 farmers in the five agricultural regions of Zimbabwe and was supported by the use of the Internet. The findings reveal that some farmers have adopted IoT in Zimbabwe, others are still to adopt such technology and some are not aware of the technology. IoT's benefits to Zimbabwean farmers are immense in that it improves food security, water preservation, and farm management. However, for most farmers to benefit from IoT, more awareness campaigns should be carried out and mobile and fixed Internet connectivity improved in some of the areas.

KEYWORDS

Internet of Things, Adoption, Smart Agriculture, Activity Theory, Covid-19.

1. INTRODUCTION

In Zimbabwe agriculture forms the backbone of the economy by contributing approximately 17% of the Gross Domestic Product (GDP). Farming activities generate an income for about 60 -70% of the population [1]. The Zimbabwean agricultural sector is composed of crop production, animal production, and forestry (tree plantations) [2]. This sector has seen a decline in food production, deforestation resulting in the country importing major food items such as maize, wheat, and soya beans. The introduction of technology in agriculture has boosted food production in some of the developed countries especially in the United States of America (USA) and other developing countries [3]. While most African countries are still facing difficulties in food security, later alone the adoption of technology in the agricultural sector remains a challenge.

In1999 to 2000, the Zimbabwean agricultural sector did undergo agrarian reforms to equitably share land which had been caused by colonial imbalance. Most of the farms were subdivided so that more farmers would be accommodated. The farms were previously described as communal, resettlement, small-scale commercial, and large-scale commercial farms. During the agrarian reforms, the farms were modelled along with A1 and A2 models. A1 model farms are divided into small plots where a number of villagers are apportioned 5 hectares of arable land with

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.99-106, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111208

communal grazing. A2 model comprises of autonomous farms with commercial activities. The farming activities for both A1 and A2 models are based on the agro-ecological region[4].

Five agro-ecological regions in Zimbabwe are classified as natural regions. These regions are categorized according to the amount of rainfall, soil quality, vegetation, climatic conditions among other factors. Zimbabwe's rainfall pattern ranges from 550 to 900 millimeters across the five regions. Most of the Zimbabwean farmers rely on rainfall for crop farming and some of the A1 and A2 farms largely rely on irrigation [5]. Zimbabwe has experienced food shortages due to droughts, storms, and floods. This has forced the country to import food for the past decade [6]. There is a growing need for Zimbabwean farmers to utilize technology to improve food security.

2. LITERATURE REVIEW

Internet of Things is defined as a network of interconnected devicessuch as sensors and communication networksconnected through the internet to transfer information without human intervention [7]. IoT has managed to change the traditional method of farming, by aiding farmers with the use of technology. This has transformed the agricultural sector from precision farming into smart farming [8]. A farmer can monitor their field with the use of sensors like dielectric soil moisture sensors. The sensors give an opportunity for a farmer to plan watering times and areas that need to be irrigated the most. Sensors can also be used to monitor and alert the farmer on the movement of pests in the field. IoT allows farmers to remotely control farm activities, processing, and logistic operations by the use of sensors and actuators, e.g. it allows for accuracy in the application of pesticides and fertilizers or robots for automatic weeding. IoT can be used to monitor food quality during transportation by remotely accessing and controlling the geographic location and conditions of shipments and products.

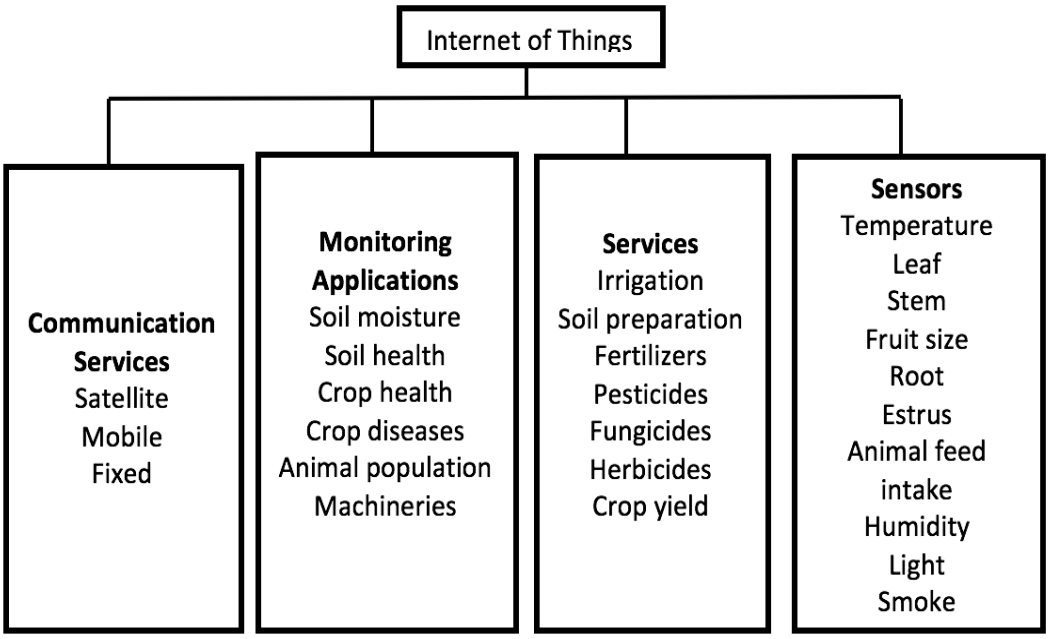


Figure 1. Illustration of Internet of Things in agriculture

Figure 1 illustrates the application of the Internet of things in agriculture which is composed of four elements, communication services, monitoring applications, services, and sensors. Communication services include network services for Internet data that can be offered through

satellite, mobile fixed networks [9]. Data transmission of IoT devices varies and can be supported with 2G-5G cellular networks. Internet-linked devices enable farmers to collect and exchange data without human involvement. The monitoring applications can be used to monitor soil moisture, soil health, crop health, crop diseases, and animal population [10]. Machinery such as combine harvesters, tractors, irrigation equipment, and drones can be fitted with sensors [11], for example, Hello Tractor developed a low-cost monitoring device that can be used to monitor the condition of the tractor [12]. IoT can be applied through an agricultural drone which is a relatively inexpensive device fitted with a mechanism that provides farmers with information about the status of the crops which can result in an increase in yields and reduce crop damage.The drone can also be used to track and monitor the movement of animals and check if there is any danger being paused in their area [13]. IoT can also be used with irrigation equipment where water usage can be monitored. The services include the detection of soil nutrients and the amount of fertilizer required. The services for IoT in agriculture vary from crop yield to, detection of pests and herbs affecting the growth of the crop. Sensor devices play a pivotal role in the collection of data about the status of the land, crop, or animal, for example, the devices can be used to determine fruit size, moisture, or nutrient content [14].

With the adoption of IoT, farmers will be able to control the internal processes and thereby decrease production risks. The availability of data allows farmers to foresee the output of production and allows for better planning especially crop management and product distribution. With enhanced control over overproduction, waste levels can be reduced and costs can be more effectively managed. Knowledge about any anomalies or challenges in the rate of crop growth or the health of livestock allows farmers to mitigate the risk of diminished yield or even crop failure [15],.

To implement an effective and successful IoT solution in agriculture three factors should be considered:i) allowing for real-time collection and presentation of data, ii)providing a solution that is low-powered, easy to install, and cost-effective and iii) provide a solution that can be remotely accessed globally and not restricted to the operator or local networks [16]. Farmers are more worried about farm management and increase in production, rather than being bogged down about technical and costs issues of IoT.

Current agricultural trends have seen the adoption of novel strategies of crop production such as greenhouses, hydroponics, vertical farming, and phenotyping to increase crop yield [14]. Crop production in greenhouses is done in a controlled environment which allows for seasoned and unseasoned crops to be grown anywhere at any time. Wireless communication, mobile devices, and other Internet devices are used in the greenhouse to monitor humidity, temperature, light, and pressure. Hydroponics allows farmers to grow seasonal and unseasonal crops in water under controlled conditions without a soil medium and the nutrients are applied through the irrigation system. Wireless devices connected through the Internet are used to monitor the water level, nutrients, and fertilizers used for crop production. Vertical farming allows farmers to grow crops in a controlled environment on a small piece of land. This type of farming is commonly used in Japan [17]. The use of IoT in vertical farming permits the control of moisture and groundwater using computers or cellular devices such as tablets and smartphones. Phenotyping *"is an advanced genetic engineering technique and biotechnology which correlates the genetic sequences of crops for agronomical and physiological aspects"* [18]. In this approach, IoT is used to determine and analyze the characteristics of genetic engineering and biotechnology of the crops [7].

IoT is also being used to improve the sustenance of food production in aquaculture. Aquaculture is an agricultural activity where farmers focus on producing fish, water plants, and diverse oceanic organisms [19]. Devices can be used to monitor the water, oxygen and nutrients levels

and transmit this data through the Internet. Aquaponics is a sustainable agriculture in a symbiotic environment by combining aquaculture and hydroponics [20]. The water system should flow on the planting medium periodically to ensure the plants get the nutrients, while the water can be filtered properly by the medium.

IoT can be used with cloud computing which resolves some of the limitations of the devices and sensors, by providing storage solutions and computing power for analysis. Cloud computing also offers farmers an opportunity to obtain valuable information about markets, especially seeds, fertilizers, equipment, and farming methods. Cloud computing can also facilitate the use of Big Data analytical tools for farmers [21] [22].

Most farmers in Zimbabwe spend their time physically monitoring and understanding farming activities while modern agricultural activities require better farming management techniques through the adoption of technology. In developed countries, farmers have adopted intelligent farming systems to improve agricultural activities [23]. Zimbabwe has been affected by severe droughts in the past two decades which has resulted in lower food production. Does the adoption of technology (IoT) efficiently and effectively help improve agricultural production in Zimbabwe? This paper investigated how the adoption of IoT for smart agriculture in Zimbabwe will help improve food security.

The theoretical framework of this research is based on the activity theory which seeks to describe the socio-technical activities in agriculture. The activity theoryhelped the researchers in understanding the processes, actors involved in farming activities, and how the adoption of IoT will benefitZimbabwean farmers.

3. RESEARCH METHODOLOGY

The mixed methodology was used in this research involving the collection of qualitative data throughinterviews and documentation. 50 farmers from A2 model farms in the five agro-ecological regions were interviewed online due to the Covid- 19 restrictions. The Covid – 19 restrictions in Zimbabwe limited the movement of people to curb the transmission of the disease. Data was also gathered from the Internet about the adoption of IoT in the Zimbabwean agricultural sector. Farmers who had access to the Internet and Social media platforms were identified through purposive sampling. The six tenets of activity theory: the objective of the IoT, the actors involved, the agricultural community, the technology (IoT), the division of activities among the actors in the agricultural system, and regulations governing the use of IoT in Zimbabwe were used as guidelines for the research. Data was then analyzed and categorized into three clusters, i) farmers who had adopted IoT, ii) farmers knowledgeable about IoTbut had not adopted such technology, and iii) those who were not aware of such technology.

4. FINDINGS

From the interviews held online, three clusters emerged, farmers who had adopted IoT, farmers knowledgeable about IoT but still to adopt such technology, and others who were not aware of such technology.

4.1. Farmers Who Have Adopted IoT in Agriculture

During the interviews, some of the farmers revealed that they had adopted IoT and the benefits were quite enormous. The farms that had adopted IoT, were able to monitor soil nutrients, moisture, water usage, temperature, humidity, light, weed, and pests. Some of the farms had

sensors in their greenhouses to monitor the environmental parameters for example the sensors were able to monitor temperatures, humidity, soil nutrients, and light. In one of the farms, the farmer installed global tracking devices on some of the bull cattle. This assistedthe farmer in animal management i.e. ability to locate the animals if they had been lost or stolen.

The following benefits were highlighted on the farms that were adopting IoT in Zimbabwe: enhanced decision making, savings in electricity, preservation of water, better yields, and reduced labour. The farmers indicated that they were able to monitor their fields or animals remotely and could make faster decisions especially if they had challenges on the farm. The farmers also stated that electricity was saved due to constant monitoring of the moisture content of the soil rather than physically checking the wetness of the ground. The farmers also revealed that water usage was reduced because only areas that needed to be irrigated would get the required amount of water. There was an improvement in the yields as farmers were able to monitor the growth of their plants especially the soil nutrients and other adverse weather conditions. Labour costs were also reduced as the farmers did not have to send someone to the field to physically check the temperatures, moisture, or nutritious content. This data would be remotely transmitted to the farmer.

4.2. Farmers Knowledgeable about IoT

During the interviews, some farmers were knowledgeable about IoT, but are still adopting the technology. The reasons that were given by the farmers for not adopting such technology were cost, lack of proper infrastructure, poor internet connectivity, and the requisite skill to adopt such systems. The farmers stated that adopting such technology using the existing mobile or fixed networks had challenges during access, uploading, and downloading data as services were poor and not accessible in some other areas. The other alternative which is satellite services were said to be costly in Zimbabwe.

4.3. Farmers still to adopt IoT in agriculture

The majority of the farmers interviewed were not aware of the IoT technology and its benefits and it was their first time to be introduced to such technology. Some of the farmers who did not know about the existence of such technology revealed that they were eager to embrace this innovative technology. But, some of the farmers, although made aware of IoT in agriculture through this interview said they would not adopt such technology.

5. DISCUSSION AND ANALYSIS

In some of the farms in Zimbabwe, IoT has changed the traditional methods of farming to Smart Farming. Farming activities have formed a smart web of interoperable farm objects. With the adoption of IoT, farm management is integrated by real-time sensing and monitoring, smart analysis and planning, and smart control of all relevant farm processes [24].

IoT can be adopted by monitoring the seeds, plantation, harvesting, and quality of the products during the whole cycle of crop production. Many benefits come with the adoption of IoT by Zimbabwean farmers such as remote monitoring of farming activities and enhancing the decision-making process as evidenced in section 4.1. Although the benefits of IoT in agriculture are enormous, some farmers are not knowledgeable about such technology in Zimbabwe. This is supported by [25] who state that the lower levels of education in technologies in developing countries affect the technological acceptance by farmers as they are more comfortable with traditional methods of farming as compared to the modern techniques. [25] further alludes that

most farmers pride themselves in crop and animal production, the terrain and soil quality of their land, and are reluctant to adopt high-tech monitoring technology.

A developing country like South Africa is implementing IoT in agriculture examples are in the wine industry which monitors the whole cycle of growing grapes up to the level of wine production. IoT is also used in other crop production such as potatoes, maize production water, and livestock monitoring [26].

Real-time communication plays an important role in the adoption of technology in advanced agriculture. Communication enhances faster decision making thereby enabling farmers to manage farm activities effectively [27]. Network service providers should provide services that will benefit farmers in the adoption of IoT. The advancement of IoT requires more bandwidth and the cost of data from the network service providers should be affordable to farmers in Zimbabwe. However, when approached effectively, the adoption of IoT and more broadly information and communication technology in developing countries like Zimbabwe can contribute towards food security.

By using smart agriculture technology, Zimbabwean farmers will gain better control of farming activities such as the rearing of livestock and growing crops, bringing about massive efficiencies of scale, cutting costs, and helping save scarce resources such as water.As Zimbabwe is often affected by droughts, IoT will allow the country to preserve water. With the adoption of IoT, Zimbabwe will be able to provide smart solutions in agriculture.

6. CONCLUSIONS

The use of IoT has immense benefits to Zimbabwean food security as farmers will be able to make faster decisions thereby boosting agricultural productivity. IoT will enable framers to monitor soil nutrients, environmental parameters, water usage and enabling farmers to be well informed about agricultural activities in their fields regardless of geographic area.

However, it is important to note some of the farmers are not aware of IoT and some of the agricultural regions have poor Internetconnectivity making it difficult to adopt such technology. It is recommended that awareness campaigns should be conducted for the adaptation of IoT by farmers in Zimbabwe. In order for Zimbabwean farmers to also benefit from the adoption of IoT, communication infrastructure coverage and speed of the Internet should be improved.

ACKNOWLEDGEMENTS

This paper is inspired by the research that has been done about IoT in the Zimbabwean agriculture sector.

REFERENCES

[1] FAO, “Zimbabwe at a glance: Food and Agriculture Organization of the United Nations,” *FAO Zimbabwe*. 2018.
[2] Ministry of Agriculture, “Comprehensive Agricultural Policy Framework,” 2012.
[3] OECD, “Adoption of Technologies for Sustainable Farming Systems,” in *Wageningen Workshop Proceedings*, 2001, p. 147.
[4] S. Moyo, “Review of African Political Economy,” *JSTOR*, vol. 38, no. 128, pp. 257–276, 2011.
[5] FAO, *Fertilizer use by crop in Zimbabwe*. 2004.
[6] K. Nyikahadzoi, “Drought Hazard Risk and Humanitarian Impact Analysis and Inventorisation of Forecast Models in Zimbabwe,” 2021.
[7] N. Khan, R. L. Ray, G. R. Sargani, M. Ihtisham, M. Khayyam, and S. Ismail, “Current progress and

future prospects of agriculture technology: Gateway to sustainable agriculture," *MDPI*, vol. 13, no. 9, pp. 1–31, 2021.

[8] V. Saiz-Rubio and F. Rovira-Más, "From smart farming towards agriculture 5.0: A review on crop data management," *MDPI*, vol. 10, no. 2, 2020.

[9] J. E. Sierra, B. Medina, and J. C. Vesga, "Management system in intelligent agriculture based on Internet of Things," *Espacios*, vol. 39, no. 8, 2018.

[10] H. M. Jawad, R. Nordin, S. K. Gharghan, A. M. Jawad, and M. Ismail, "Energy-efficient wireless sensor networks for precision agriculture: A review," *MDPI*, vol. 17, no. 8, 2017.

[11] FAO, "Agriculture 4.0: Start Agricultural robotics and automated equipment for sustainable crop production," 2020.

[12] Hello Tractor, "Hello Tractor, Break Ground, Drive Change."

[13] R. Giacomo and G. David, *E-Argriculture in action: Drones for agriculture*. 2018.

[14] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E. H. M. Aggoune, "Internet-of-Things (IoT)-based smart agriculture: Toward making the fields talk," *IEEE Access*, vol. 7, pp. 129551–129583, 2019.

[15] R. A. Acharige, M. N. Halgamuge, H. A. H. S. Wirasagoda, and A. Syed, "Adoption of the Internet of Things (IoT) in Agriculture and Smart Farming towards Urban Greening : A Review," *Int. J. Adv. Comput. Sci. Appl.*, no. April, pp. 10–28, 2019.

[16] C. Verdouw, S. Wolfert, and B. Tekinerdogan, "Internet of things in agriculture," *CAB Rev. Perspect. Agric. Vet. Sci. Nutr. Nat. Resour.*, vol. 11, no. October 2017, 2016.

[17] K. Benke and B. Tomkins, "Future food-production systems: vertical farming and controlled-environment agriculture," *Sustain. Sci. Pract. Policy*, vol. 13, no. 1, pp. 13–26, Jan. 2017.

[18] N. Khan, R. L. Ray, G. R. Sargani, M. Ihtisham, M. Khayyam, and S. Ismail, "Current progress and future prospects of agriculture technology: Gateway to sustainable agriculture," *Sustain.*, vol. 13, no. 9, 2021.

[19] U. Acar *et al.*, "Designing An IoT cloud solution for aquaculture," *Glob. IoT Summit, GIoTS 2019 - Proc.*, no. June, 2019.

[20] C. Li *et al.*, "Prospect of aquaponics for the sustainable development of food production in urban," *Chem. Eng. Trans.*, vol. 63, no. August, pp. 475–480, 2018.

[21] G. Vitali, M. Francia, and M. Golfarelli, "Crop Management with the IoT : An Interdisciplinary Survey," *MDPI*, pp. 1–18, 2021.

[22] S. Nandhini, S. Bhrathi, D. D. Goud, and K. P. Krishna, "Smart Agriculture IOT with Cloud Computing, Fog Computing and Edge Computing," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 2, pp. 3578–3582, 2019.

[23] A. Walter, R. Finger, R. Huber, and N. Buchmann, "Opinion: Smart farming is key to developing sustainable agriculture," in *Proceedings of the National Academy of Sciences*, 2017, vol. 114, no. 24, pp. 6148–6150.

[24] N. Dlodlo and J. Kalezhi, "The internet of things in agriculture for sustainable rural development," *Proc. 2015 Int. Conf. Emerg. Trends Networks Comput. Commun. ETNCC 2015*, pp. 13–18, 2015.

[25] E. Botha, R. Malekian, and O. E. Ijiga, "IoT in Agriculture: Enhanced Throughput in South African Farming Applications," *2019 IEEE 2nd Wirel. Africa Conf. WAC 2019 - Proc.*, no. September, pp. 1–5, 2019.

[26] P. Aguera, N. Berglund, T. Chinembiri, A. Comninos, A. Gillwald, and N. Govan-Vassen, "Paving the way towards digitalising agriculture in South Africa," no. June, pp. 1–42, 2020.

[27] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Sci. Direct*, vol. 58, no. 4, pp. 431–440, 2015.

AUTHORS

Tsitsi Zengeya is a Computer Science lecturer at Great Zimbabwe University. She holds an Msc degree in Computer Science. Areas of research interest are: Artificial Intelligence, Data Science & Ontologies.

Dr. Paul Sambo is a lecturer at Great Zimbabwe University's Computer Science department. He has wealth of experience in the Information Communication Technology industry. He holds a PhD in Information Systems and his research areas are: Information systems, Data Communication, Artificial Intelligence and Software Engineering.

Nyasha Mabika is a lecturer at Great Zimbabwe University, Department of Livestock, Wildlife and Fisheries. He has vast experience in biological techniques and fishery sciences. He is a PhD holder with an interest in fish health and bio monitoring.

LYRICS TO MUSIC GENERATOR: STATISTICAL APPROACH

V.N Aditya Datta Chivukula[1], Abhiram Reddy Cholleti[2] and Rakesh Chandra Balabantaray[1]

[1]Department of Computer Science and Engineering
[2]Department of Electronics and Telecom Engineering
International Institute of Information Technology, Bhubaneswar, Odisha, India

ABSTRACT

Natural Language Processing is in growing demand with recent developments. This Generator model is one such example of a music generation system conditioned on lyrics. The model proposed has been tested on songs having lyrics written only in English, but the idea can be generalized to various languages. This paper's objective is to mainly explain how one can create a music generator using statistical machine learning methods. This paper also explains how effectively outputs can be formulated, which are the music signals as they are million sized over a short period frame. The parameters mentioned in the paper only serve an explanatory purpose. This paper discusses the effective statistical formulation of output thereby decreasing the vast amount of estimation of output parameters, and how to reconstruct the audio signals from predicted parameters by using 'phase-shift algorithm'.

KEYWORDS

Audio Signal Processing, Statistical Machine Learning.

1. INTRODUCTION

Natural language processing algorithms are very helpful in understanding human-computer interaction effectively. But it becomes necessary to try different approaches in managing the same problem as it enhances our understanding and may spark up many brilliant ideas in different fields. This makes overall research in the field effective. In recent times generator has become one of the most important fields of study because of many reasons including its capability in understanding human-like thinking in specific fields like music, literature, and so on. This also helps us in increasing the capability of capturing the complex features and reciprocating them till the output. Hence, any generator algorithm can become very important in understanding various applications and details on which the basic functionalities are dependent. Hence, it is definitely a growing field in terms of research and application of which we tried to propose a new approach towards understanding it.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.107-119, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111209

2. LITERATURE REVIEW

Lyrics to music generators is a certain application that is in its nascent stage and yet has overwhelming demand in the research application field. Currently, the most optimum results are published by jukebox [1]which is a project by OpenAI. In this work they used deep learning techniques where they actually fed the raw data without using any extraction techniques to separate voice from the original song, instead they used raw audio to train the model which spits out raw audio in return, where they had to use symbolic music for better results. They have concentrated more on working with VQ-VAE [2] model to compress raw audio data and then used autoregressive transformers [3]. The model actually learn the compressed audio generation which is generated in the form of discrete codes and it makes the process computationally expensive in the data pre-processing stage itself and has a risk of overfitting. They have used the attention model [4] mechanism extensively to learn the interdependencies between the lyrics and capture the similarities which are very risky as the dataset at the current stage is very less. The dataset the main model is learning is also an output of autoregressive transformers which is actually decreasing the originality of the information itself which may result in content tampering as a result of overfitting or underfitting. We propose a completely different approach and well-established facts and theory which is completely based on statistics during pre-processing of data and preserving the original information intact. We also propose the decoding part in a much simpler sense using a simple fact of superposition which works almost in every case without actually using deep learning approaches. Our propose here is a purely statistical approach right from encoding, learning parameters for which we used a regression algorithm [5], and finally, for the decoding part we have come up with a new algorithm that effectively works in the majority of cases.

3. PROPOSED METHOD

The different components of the proposed method along with the results are mentioned below in various subsections.

3.1. Voice Separation

We know that music is recorded on both left and right channels, and the vocals which are present in the music are mixed with different instruments and it becomes hard to remove them. We used Fast Fourier Transformation [6] to make a comparison between the left and right channels with the vocal and non-vocal part of that song. On experimentation, we found that the amplitude corresponding to vocals are present when the magnitude of FFT of the left channel sample is greater than 100 and that of the right channel sample is less than 400, if the vocals are in a low pitch and have no chorus in the background. Hence, if the value lies in that range we assigned the FFT value to be 0 for that sample. After taking the inverse Fourier transform of both the channels, we found that the voice was decent when subtracting those channels from each other but the noise was developed in the signal. On experimenting with an instrumental track, we compared the frequency domain of the resultant signal and instrumental signal, and found that noise ranges from 900Hz to 14000Hz (approx.). To remove noise we used the band-stop filter. Figure 1, represents the extracted signal in the time domain while Figure 2, represents the extracted signal in the frequency domain, we can observe that signals contributing inthe frequency range 900Hz to 14 KHz were zero.

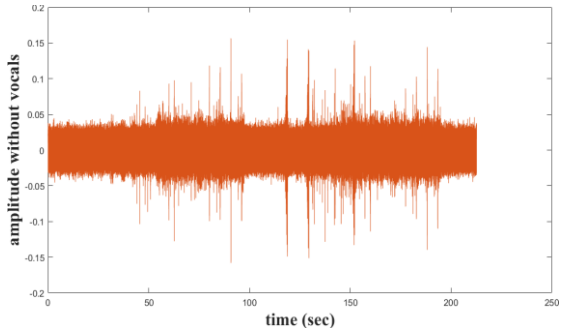


Figure 1. Amplitude Vs Time (Extracted Signal)

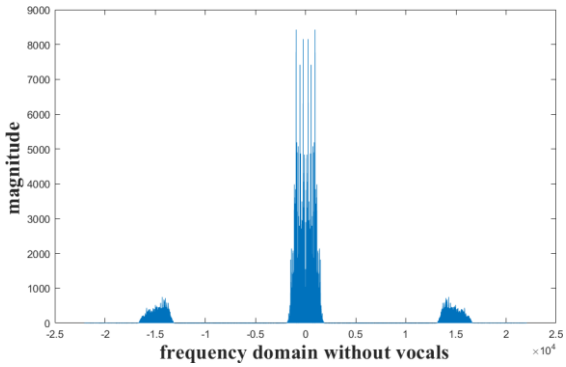


Figure 2. Extracted Signal in the frequency domain

On experimentation, we found that the angle between the music and its instrumental track is between 85-90 degrees. Figure 3, depicts the process of voice separation.

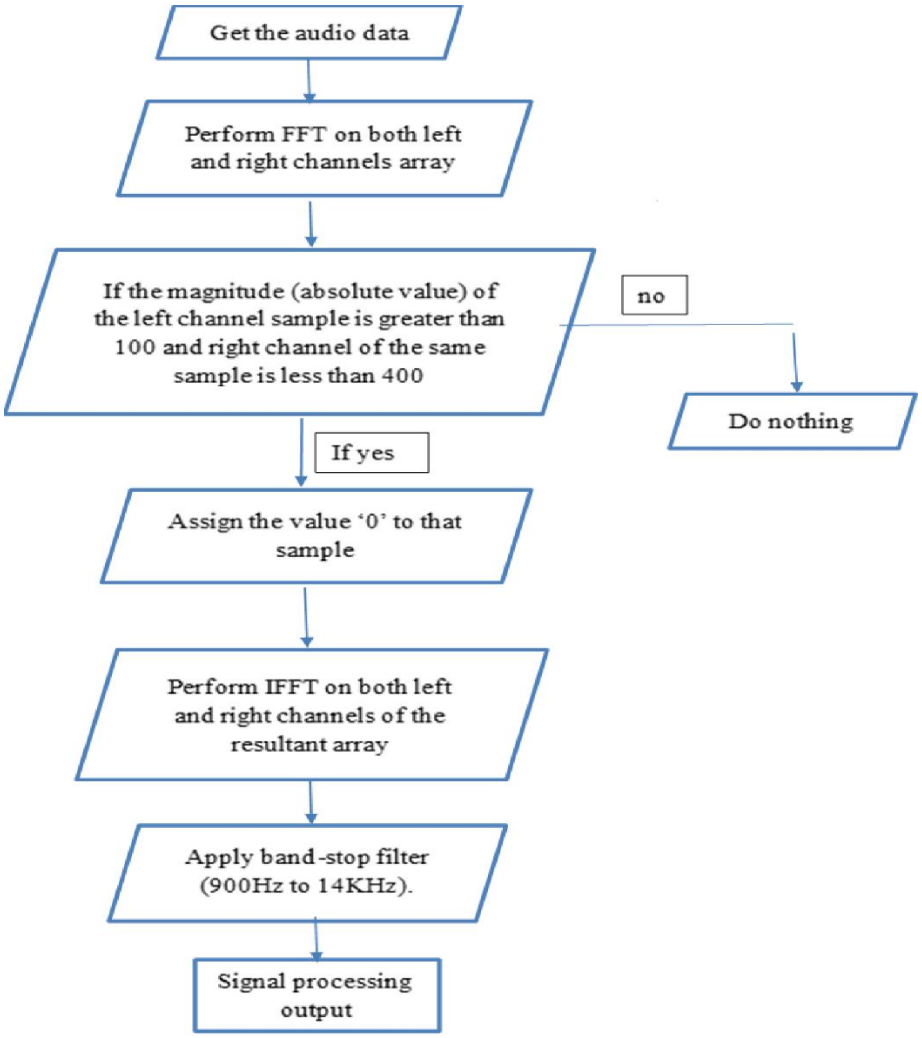


Figure 3. Flowchart of voice separation

3.2. Gaussian Representation

This subsection simply discusses the normalization of the audio array, One thing to be noted is that before normalization of the audio array, note down the mean and standard deviation of the array and label these parameters as O^1 and O^2 (O^1 and O^2 are two out of six output parameters we need to predict for test set). From the remaining four parameters, two parameters are the minimum and maximum value of the resultant array we get after performing normalization, which we are going to store for each sample of the audio dataset. Now, one has to just divide the absolute difference between the maximum and minimum value of the normalized array into 1000 intervals or more depending upon the range of minimum to a maximum value, while we were experimenting. We just selected 1000 intervals at random (as the number of intervals can be treated as a hyperparameter, one can always experiment with different values). Now, count how many values in the normalized array are falling under this specific interval and divide this count with the dimension or size of the normalized array. In this way, one can get the probabilities of each interval and if a graph is plotted taking the intervals on X-axis and corresponding probabilities on Y-axis, then, we get the approximate Gaussian curve as gaussian formula suggests one can get the Gaussian distribution [7] of n samples of data by converting the given data as follows

$$X_n^{|} = (X - \mu)/ (\sigma/ \sqrt{n}) \qquad (1)$$

where μ is the mean of the sample, σ is the standard deviation and 'n' is the sample size. For the implementation we have taken n=1 as there is no more than one song from the same distribution because one can observe each song's mean and standard deviation is different but, we will get an approximate gaussian as in Figure 4.

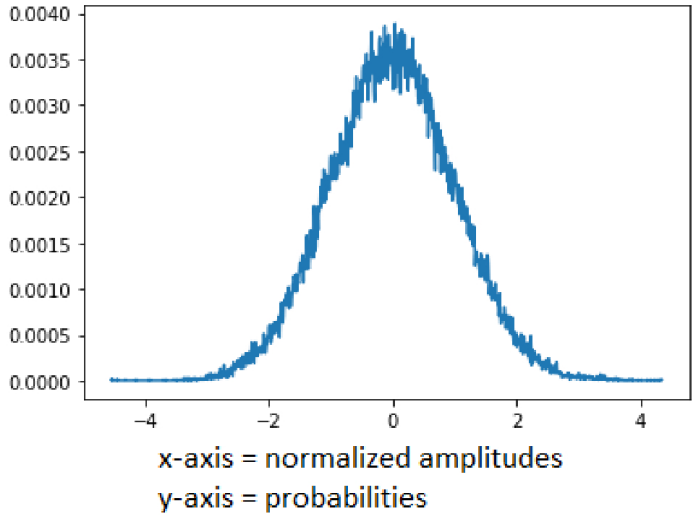


Figure 4. Depiction of approximate Gaussian

One, important feature or benefit one can get with this trick for music analysis is the enormous compression of data for the processing which decreases the data by (size of array-number of intervals) *100%. To understand this compression, if the size of the array is 10000000 and the number of intervals is 10000 then the compression, we have got is 1000 times less than the original which is, reducing the original to 0.1% of its size with minimal loss in information.

In Figure 5 one can see the values reconstructed by replacing every value of an interval with the mean of interval on the X-axis and original normalized array values on the Y-axis. The normalized array values and reconstructed array values are both in ascending order, so that, the graph represents only how close the reconstructed values are to the true normalized values without the arrangement of amplitudes which will be discussed later as to how to arrange these values in order close to the original normalized arrangement of amplitudes. Theentireprocess can be understood with the flowchart as shown in Figure 6.

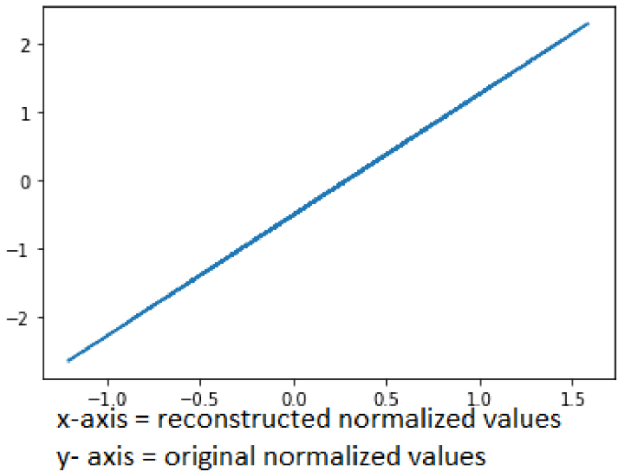


Figure 5. Depiction of similarity between Reconstructed: values and original normalized

3.3. Input Formulation

For designing the input, we have taken the lyrics and pre-processed it by performing sentence tokenization, word tokenization keeping in mind the chunking of characters which have expressional characters such as, '?', '!' etc, as they really influence the music. Because we assume that expressions are very much necessary in creating the music and one can never neglect the expression or feel or emotion in the lyrics. Then, we have taken the limit of 700 tokens for each song for applying the machine learning applications in the future. Every song is processed in the same way and made into a dataset of 100 songs each containing 700 tokens. Then, this dataset is passed on to the word embedding section where the word2vec library is used for converting the tokens into word vectors using a skip-gram model [8]. Then we get each token in a vector form, but we need a single value to represent the token but not a vector, and hence, one can just take the magnitude of the vector and embed the word token with its magnitude. Hence, the dataset can be reformulated with magnitudes of each token, if for a song the number of tokens is less than 700 then we can pad the remaining with zeros. Thus, the above input formulation can be completed. An important note is that the above-mentioned values are just for explanatory purpose and one can always try different deep learning and machine learning models to actually map the lyrics to the required output parameters. The value 700 is used here because there was no song that has more than 700 tokens, hence, this value acts as an upper bound for padding of sequences for input part training data. In Figure 7 there is a flowchart depicting the procedure followed for input formulation.

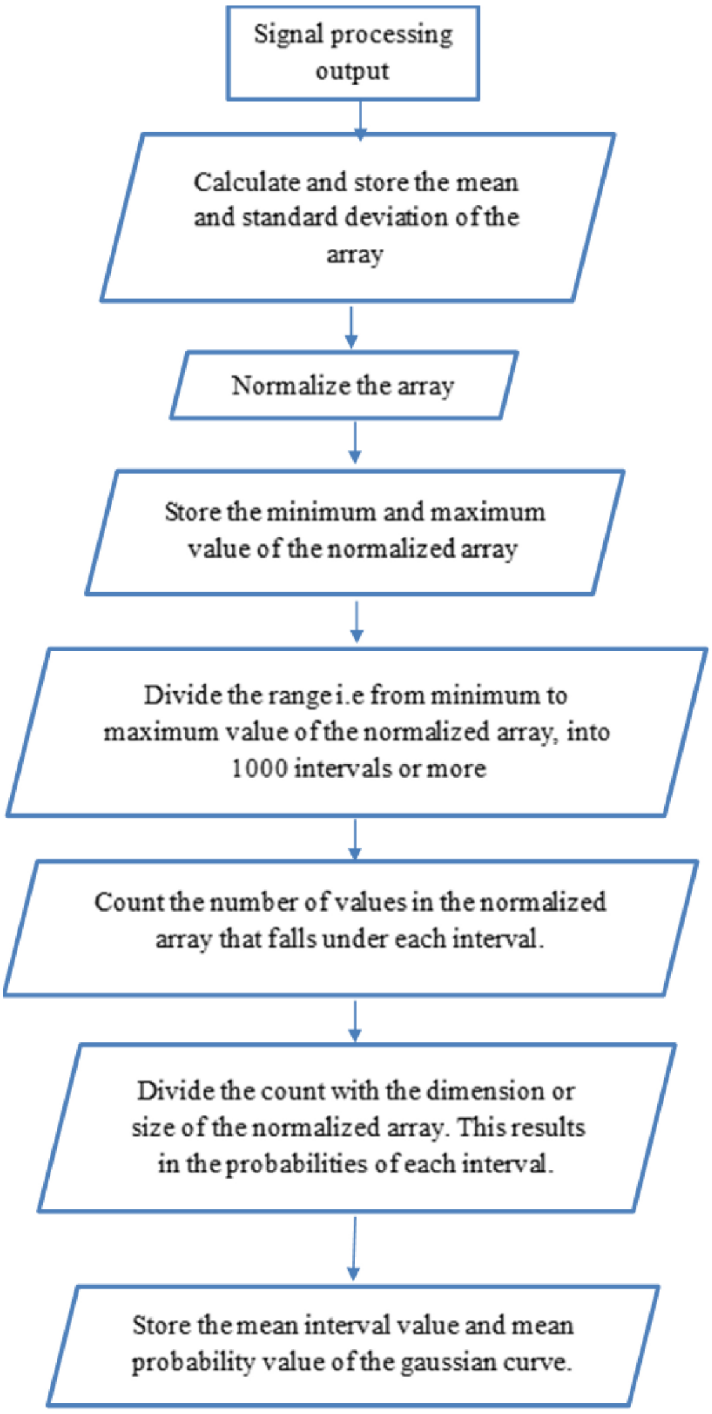


Figure 6. Flowchart of gaussian representation

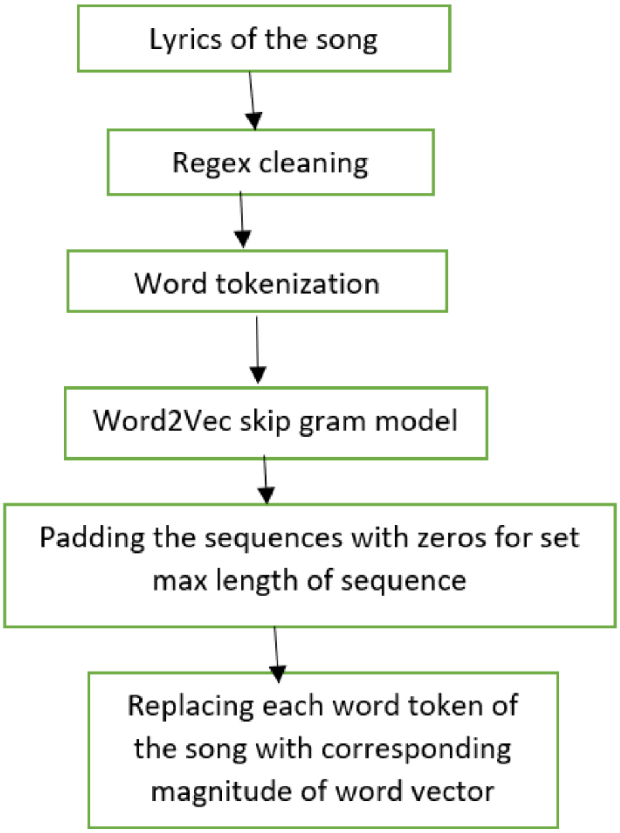


Figure 7. Flowchart depicting input formulation of the dataset

3.4. Output Formulation

As discussed in section 3.2, four out of six output parameters are mentioned and now the remaining two parameters are nothing but the mean frequency and mean probability, which implies, the total six parameters are:

Original array mean, original array standard deviation, normalized array minimum, normalized array maximum, gaussian mean interval, and gaussian mean probability. Each of the six output parameters is predicted by training them separately with six copies of linear regression models on the same input dataset as represented in Figure 8.

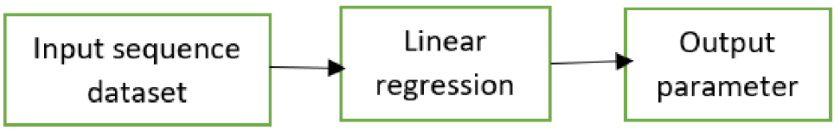


Figure 8. Block diagram depicting the training of machine learning algorithm

Once, one has obtained the parameters, first, they can take the mean interval and mean probability and calculate the standard deviation of predicted gaussian, then, take the mean and standard deviation of the predicted gaussian to actually construct the Gaussian curve and fix the X-axis range according to the minimum normalized array value and maximum value. Now, One can divide the range into a number of intervals they used while training for the experiment, we took 1000 intervals and calculated the average probability of each interval and store it in a 1000

sized array of probabilities. Finally, we just need to take an average interval value of 1000 intervals by which we get the average frequency array and repeat each value of frequency array into a new array of a size equivalent to normalized array size according to the corresponding probability of occurrence i.e if the first value in the probability array is 0.01 and average first interval value is -4.1 and also if the size of the normalized array used for training is 100000 array then the first 1000(0.01 * 100000) values of predicted normalized array should have -4.1 as its value and one has to repeat the same for remaining 999 intervals and corresponding probabilities. In this way, one can reconstruct the normalized predicted array, from this array taking the predictions of the original array to mean and original array standard deviation values we can reconstruct the original version of the array. But this original array will be in ascending order as we travelled the gaussian from left to right intervals. To tackle the original arrangement problem, we will discuss its solution in the arrangement section.

3.5. Predicting Outputs

Given input and output, one can test the results either by taking the deep learning approach or the statistical machine learning approach. When we tried to predict six output parameters each by using a simple neural network [9], CNN [10], RNN [11], and also LSTM [12], we did not get descent results of predictions and this may be because of the fact that dataset size is less. Therefore, we went for a statistical linear regression [13] technique by training each output with only one regression model for better predictions by which we used six copies of the linear regression model and six copies if same input data to train six output parameters individually and got decent values or predictions even on the low dataset. Hence, with this method we made the entire generator working completely on statistics. Just for example, In figure 10 we can see that some part of the graph is a straight-line relationship between the predicted amplitudes and original music amplitudes both arranged in ascending order, for a randomly selected test case. Hence, we can say that for a small dataset too we are able to capture the amplitudes reasonably well. We have taken 70 songs in training and 6 songs for testing and predicted the results, out of which is a predicted gaussian in Figure 9.

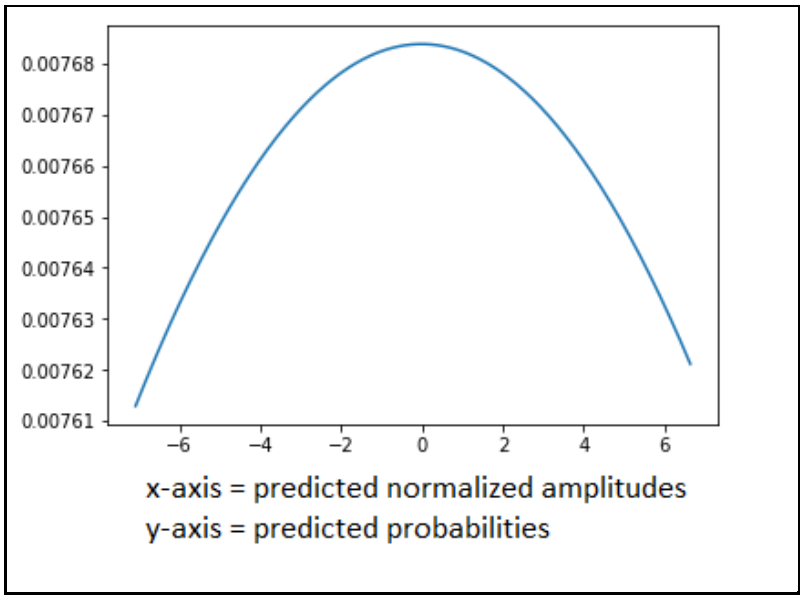


Figure 9. Plot depicting predicted normalized amplitudes (x-axis) and corresponding probabilities (y-axis)

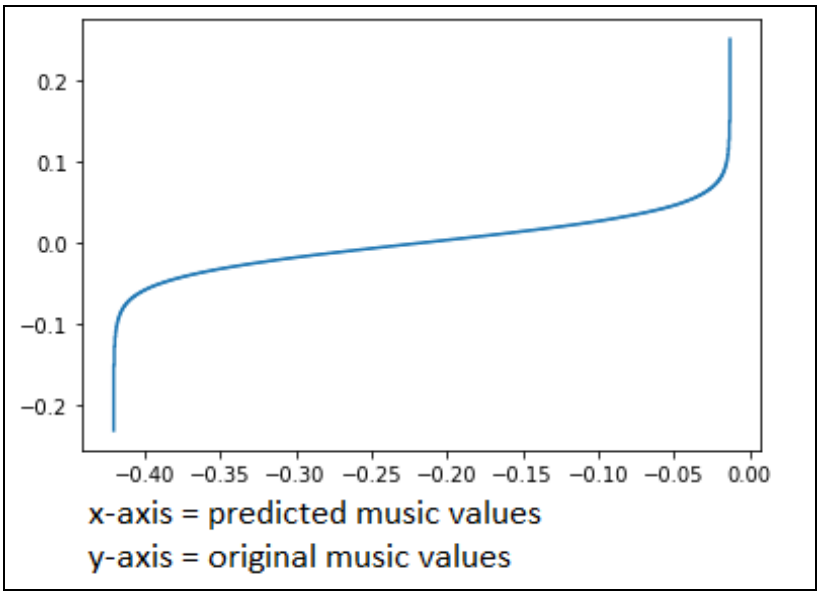


Figure 10. Depiction of similarity between original music values and predicted values both taken in ascending order to compare frequencies

3.6. Arrangements

The gaussian representation only helps us to predict the amplitudes accurately but misses in capturing the most crucial part of audio data which is the pattern of amplitudes or position at which a particular amplitude should occur in the output audio array. Because when the array is reconstructed from six output parameters, the array of amplitudes will be in ascending order but not in the desired order of amplitudes. Hence, we have arranged the amplitudes from ascending order into the desired order. On experimentation, we found that input normalized arrays are perpendicular to corresponding input normalized arrays in ascending order. So, the task finally is to rotate an n-dimensional array or vector by 90 degrees. To do this, one can follow the most obvious approach which is:

PERMUTATION METHOD: This method is very simple, just take all permutations of the array and check which permutation is perpendicular to the ascending order version of the array, but, the complexity of this approach is O (n!) which is very bad, especially dealing with music applications where array size is in millions.

3.7. Phase Shift Algorithm

This is a new algorithm to rotate a vector by some angle if that angle is possible with given values of the array. For our case, the angle is 90 degrees. This algorithm tries to find out the permutation of the vector which is 90 degrees to the ascending order array and if a 90-degree combination is not possible it gives the combination which is close to 90 degrees.

If predicted_array = [1, 2, 3] and we are required to find a near 90-degree combination. Hence, we follow the following steps:

Take int = [1] and fix its order of values.

Add next dimension value of predicted_array to init and calculate all combination possible keeping order of init fixed i.e,

Combinations = [[1,2],[2,1]]

let, theta = 0(degrees)

Calculate the angles between each combination vector and predicted_array's first two dimensions
A1 = $\cos^{-1}$(dot product ([1,2],[1,2]) = 0 degrees, 0 is not greater than 0 which implies no update of theta.

A2 = $\cos^{-1}$(dot product ([2,1],[1,2]) = 37 degrees, i.e theta<A2<90, hence, theta = A2 = 37 degrees.

Update init as init = [2, 1].

Keeping the order of values in init as constant, add new dimension value from predicted_array and calculate all combinations again as in step2.

Init = [2, 1],

Combinations = [[3,2,1],[2,3,1],[2,1,3]],

theta = 37 degrees

A1 = $\cos^{-1}$(dot product ([3, 2, 1], [1, 2, 3]) = 44.41 and as theta<A1<90, hence, theta = A1 = 44.41 degrees.

A2 = $\cos^{-1}$(dot product ([2, 3, 1], [1, 2, 3]) = 38.2 and as theta>A2 and theta<90, hence, no update.

A3 = $\cos^{-1}$(dot product([2,1,3],[1,2,3]) = 21.7 and as theta>A3 and theta<90 ,
hence, no update.

Hence, init is updated to [3, 2, 1]

Finally, we can say the required vector = [3, 2, 1] which is at an angle 44.41 degrees to [1, 2, 3].
a 90-degree combination is not possible with [1,2,3], hence we got near a 90-degree combination.
Another example, when the phase shift algorithm is applied on the vector [0, 0, 1], we get the vector [1, 0, 0] as shown in Figure 11.

With this algorithm, we can do the desired rotation of an array or vector with a complexity $O(n^2)$ which is a tremendous speed up as compared to the O(n!) permutation method.

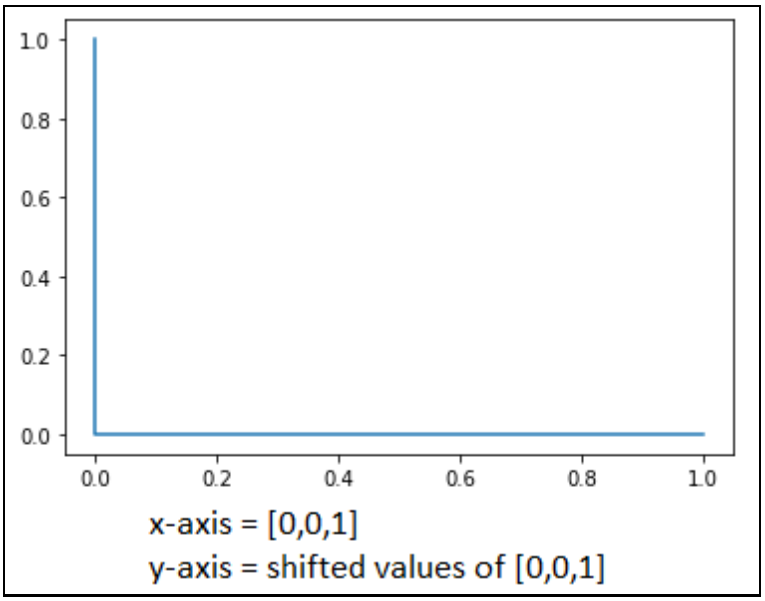


Figure 11. Graph representing a perfect shift of 90 degree for vector [0,0,1] to [1,0,0]

4. DISCUSSION

The following points are important results we got from this method:

The song and the background music of the song are mostly and approximately perpendicular to each other.

The ascending order arrangement of the music audio is mostly and approximately perpendicular to the original arrangement of the same music audio.

Gaussian distribution is used to get excellent compression of data and reduces data to (number of intervals/size of audio file array)*100% of the actual size of the data.

The phase shift algorithm rotates a given array to a given angle combination of the array if possible and if the given angle combination is not possible, it gives a near-angle combination of the array i.e an angle close to 90 degrees.

Gaussian trick enables an individual to store maximum data but at a cost of losing the sequence or arrangement of amplitudes over the time frame of the signal.

5. CONCLUSION

Generator algorithms are growing diverse as time passes incorporating better changes and many applications are also being developed on the same. The music field is one of the fields where generator applications can perform very well and do the need, by creating or generating good music and give good ideas to musicians and helping them put forward their best while composing the music. This application can serve as a resource to many in the music field and also help many lyricists and give them ideas in better and diverse ways. Generator applications can also be used in the literature field to understand how computers, if given a chance, would write, compose and speak, which is really exciting for many researchers to understand the complexities and develop better algorithms.

REFERENCES

[1] Dhariwal, Prafulla, et al. "Jukebox: A generative model for music." *arXiv preprint arXiv:2005.00341* (2020).

[2] Wang, Xin, et al. "A Vector Quantized Variational Autoencoder (VQ-VAE) Autoregressive Neural $ F_0 $ Model for Statistical Parametric Speech Synthesis." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019): 157-170.

[3] Choi, Kristy, et al. "Encoding musical style with transformer autoencoders." International Conference on Machine Learning. PMLR, 2020.

[4] Chorowski, Jan, et al. "Attention-based models for speech recognition." *arXiv preprint arXiv:1506.07503* (2015).

[5] Gupta, Swati. "A regression modeling technique on data mining." *International Journal of Computer Applications* 116.9 (2015).

[6] Cochran, William T., et al. "What is the fast Fourier transform?." *Proceedings of the IEEE* 55.10 (1967): 1664-1674.

[7] Krithikadatta, Jogikalmat. "Normal distribution." *Journal of conservative dentistry: JCD* 17.1 (2014): 96.

[8] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[9] Maind, Sonali B., and Priyanka Wankar. "Research paper on basic of artificial neural network." *International Journal on Recent and Innovation Trends in Computing and Communication* 2.1 (2014): 96-100.

[10] LeCun, Yann, and YoshuaBengio. "Convolutional networks for images, speech, and time series." *The handbook of brain theory and neural networks* 3361.10 (1995): 1995.

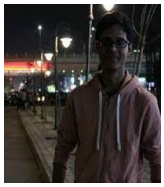
[11] Gupta, Lalit, Mark McAvoy, and James Phegley. "Classification of temporal sequences via prediction using the simple recurrent neural network." *Pattern Recognition* 33.10 (2000): 1759-1770.

[12] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[13] Goodfellow, Ian, et al. *Deep learning*. Vol. 1. No. 2. Cambridge: MIT press, 2016.

AUTHORS

V. N. Aditya Datta Chivukula is currently an undergraduate student in the Department of Computer Science and Engineering at International Institute of Information Technology, India. His area of interest in Machine learning, Deep learning, etc.



Abhiram Reddy Cholleti is currently an undergraduate student in the Department of Electronics and Telecom Engineering at International Institute of Informational Technology, India. His area of interest in antenna design, IoT, etc.

Prof. Rakesh Chandra Balabantaray research interests lie in the areas of Intelligent Systems, Data Mining, Information Retrieval, and Natural Language Processing. He teaches Information Retrieval and Data Mining, Web Mining and Natural Language Processing. He has published over 100 research papers. He has copyrights for the software he has developed. He was also the principal investigator of a project on the Development of Cross Lingual Information Access sponsored by DIT, GOI.

EFFECTIVE COMBINATION OF BERT MODEL AND CROSS-SENTENCE CONTEXTS IN ASPECT EXTRACTION

Anh Khoi Le and Truong Son Nguyen

Ho Chi Minh University of Science, National University,
Ho Chi Minh City, Vietnam

ABSTRACT

The Aspect Extraction (AE) field investigates in collecting words which are sentiment aspects in sentences and documents. Despite the pandemic, the number of products purchased online is still growing, which means that the number of product reviews and comments is also increasing rapidly, so the role of the task is gradually crucial. Extract aspects in the text is a difficult task, that requires algorithms capable of deep capturing the semantics of the text. In this work, we combine two models of the two research groups, with the first using the BERT algorithm with multiple concatenated layers and the second using the strategies to enrich the dataset by itself in the training or testing phase.
The source code is available on github.com, researchers can run it through scripts, modify it for further research also. https://github.com/leanhkhoi/AE_BERT_CROSS_SENTENCES

KEYWORDS

Sequence Labeling, Aspect Extraction, BERT, Cross-sentence.

1. INTRODUCTION

In the commercial industry, being able to capture the needs and thoughts of consumers for products is the core factor for businesses. Businesses can seize exactly what aspects of the market product users are interested in, thereby making appropriate improvements to capture customer demands. However, analyzing and extracting emotional aspects in text data such as in a sentence, a paragraph, is not lightweight works because of language complexity as well as the time-consuming dataset labeling process. So far, plenty of machine learning solutions have been proposed to carry out this task such as: POS tagger [1], Dependency parser [2], HMM [3], CRF [4], and have achieved certain results.

In recent years, deep learning algorithms have made a breakthrough and are extremely widely used as a solution to replace or supplement traditional techniques thanks to the development of hardware processing capabilities. One of the areas where deep learning has made a deep impact is NLP with the recent emergence of the Bidirectional Encoder Representations from Transformers (BERT) model [5]. Solutions based on BERT and its variants are now the state-of-the-art of many tasks in NLP. With a multi-layered architecture and trained with a huge data set, BERT is able to encode efficiently language features from syntax to semantics, providing a quality data representation layer for NLP tasks [6, 7] including Aspect Extraction.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.121-129, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111210

Besides, datasets used for Aspect Extraction task are often quite limited because of the time-consuming process of data labeling, so it is possible to use a combination of data in the same dataset come along with BERT mechanism can help to better grasp the patterns of the structure and semantics of text.

Our work is to inherit the architecture based on multiple hidden layers of BERT called parallel aggregation and hierarchical aggregation [8] in preceding research which archived quite a good performance with data enrichment techniques knows as cross-sentence [9] in another research applied for Named Entity Recognition task to create a complete pipeline so that perform Aspect Extraction task gives an even better evaluation. We also keep the configurations in our pipeline the same as those of the previous authors for experiments based comparison objective results.

2. RELATED WORKS

For unsupervised algorithms, the most popular method so far is to use a POS tagger to extract nouns or noun phrases in sentences. Stanford tagger [1] can completely do this well with an accuracy of up to 97 %, but in the problem of extracting aspects, the above accuracy does not bring much benefit because aspects in a sentence are not always a noun or a noun phrase, besides that POS tagger takes all nouns and noun phrases without any restriction according to the context of the sentence. Another approach uses relationship-based graphing, which explores relationships between emotional words and aspects based on parsing sentences into components and their dependencies. Algorithms based on this approach are referred to as Dependency parser [2], Double Propagation [10]. The weakness of these algorithms is to generate many non-aspect components. Frequency-based approaches are also considered as possible solutions. Kelledy [11], proposed a method of using POS tagger to extract all nouns and noun phrases in a sentence, then depend on frequency of occurrence of words and phrases to select the aspects. Endo [12], 2014 presented an improved method using TF-IDF in frequency calculation as well as using a syntactic pattern to remove non-aspect words and phrases.

For supervised algorithms, the two most popular traditional models for aspect extraction are HMM [3] and CRF [4]. Especially, CRF algorithm is dominating in sequential labeling problems like Aspect Extraction, the harmonious combination between BERT output and CRF is the preferred thinking in most of predicting the labels of words studies in the last 1 to 2 years.

Recently, deep learning networks have been applied to give better performance when extracting sentiment aspects, such as LSTM combined with attention mechanism [13], CNN [14]. Xu et al 2018 [15], proposed a model called DD-CNN with the idea of joining 2 embedding layers of text data, the first layer is in-domain embedding, and the second layer is out-domain embedding then feeds them into the CNN deep learning network to perform the classification task. Wang et al. 2020 [16], have introduced a mechanism to automatically concatenate pre-trained different embedding algorithms, whereby for each problem that needs embedding to form the embedding layer. For each combination, the algorithm will calculate error based on results of the training process and then compare it with other combinations to finally find out the most suitable concatenation embedding layer for the problem. This model has achieved state-of-the-art in many problems such as NER, aspect extraction (Aspect extraction), ... but has the disadvantage of expensive training time. With the arrival of BERT, a wide range of tasks in NLP such as text classification [17], summarization [18], question answering [19],… have been greatly improved in performance by using and refining this powerful BERT model. Akbar Karimi et al. 2020 [8], offers an architecture with a mix of pre-trained BERT architecture and CRF, considering Aspect Extraction problem as a sequence labeling problem which is widely applied to implement Named Entity Recognition task. The BERT architecture feature that authors use is that they are not based on only one BERT layer, usually the last layer, but they have used up to the last 4 layers of BERT

because those layers are capable of capturing more language aspects. Our research largely reuses the author's solution with some small improvements to perform Aspect Extraction task, because of the feasibility and good evaluation in experiment.

In addition to picking up the right model, we are also interested in preprocessing data to produce a quality data source. The technique used in our work is based on the ideas of Jouni Luoma et al. 2020 [9], where data will be enhanced by combining the components in itself to increase the size as well as create many interwoven semantic structures. Specifically, each data record, which is considered as a text sentence, will be joined with other data lines in the same training data set to fill the BERT window, then trained for the Aspect Extraction task according to the architecture described in the previous paragraph. Finally, the prediction of tags for each token for each sentence in the test dataset will be calculated based on the model.

3. ASPECT EXTRACTION

Given a dataset consisting of user reviews on a market product. Each line of data will contain information including aspects and respective emotions of the user. The AE's purpose is to find out exactly the aspects. For instance, *"this laptop has a good battery"*, then *"battery"* is an aspect in the sentence above. To accomplish this task, each word in the sentence will be labeled as a character in the set *{B, I, O}*, with *B* representing the starting word of the aspect, *I* representing the word belonging to one of the insides of an aspect and *O* represents the non-aspect word. The job of the algorithm is to predict each word in the sentence corresponding to each of the three characters above. This is called the sequence labeling task.

4. CROSS-SENTENCE IN CONTEXT

For each example in the dataset, it will be filled at both ends by other examples in the same dataset in a given size - BERT window size. The processed data has a fixed size (n, m) where n is dataset size and m is BERT window size. (see figure 1).

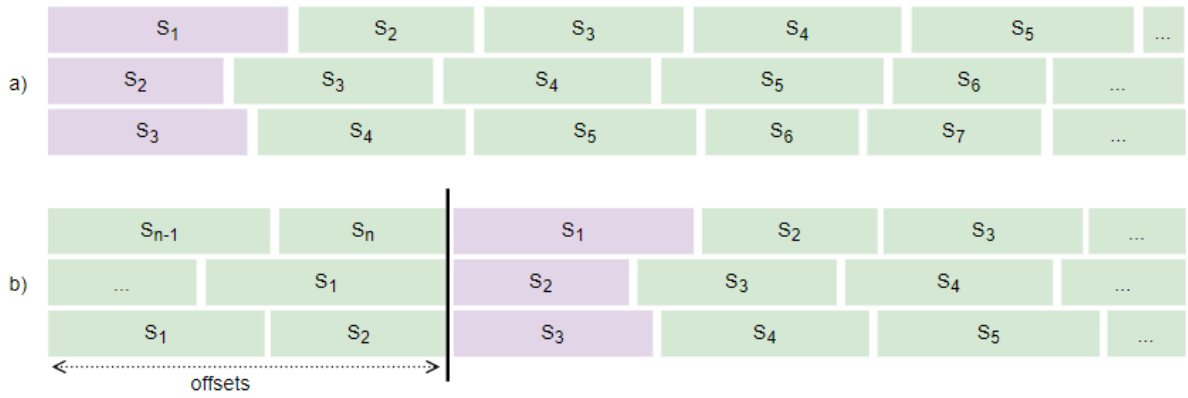


Figure 1. Illustrates two ways to create cross-sentence examples. a) The example of interest is placed at the first position in a BERT window. b) The example of interest will be placed at a position with a certain offsets value, usually 32, 64, 96,...

The data processing according to the above mechanism will create a new dataset with more semantic diversity, enabling the BERT architecture to understand more new patterns that generated from the combiner, thereby increasing the accuracy when performing sequence labeling task.

5. PROPOSAL METHOD

Our work is based entirely on the models of previous authors, bringing them together to form a complete pipeline. So we will keep the components as well as the configuration of the author's model.

Deep network models like BERT have been widely used in many problems because of their ability to understand the semantics of sentences deeply. Usually, most studies will use only last hidden layer of BERT because that is the layer that contains the most insight into the data, but authors realize that adjacent layers also store useful pieces of information [20], so they took advantage of the BERT last layers, combining them in two different ways to create the embedding layer. Besides, to be able to perform well for the sequence labeling problem, the CRF algorithm comes as an optimal solution and is widely used today. CRFs are a type of discriminative undirected probabilistic graphical model, where each data sample is predicted concerning the label of the previous data sample. CRFs are very well suited to sequence labeling tasks and Aspect Extraction can be considered as a labeled sequences problem.

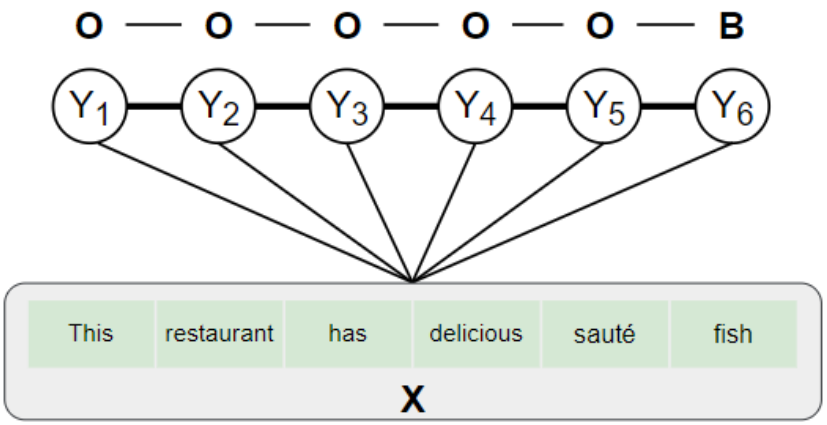


Figure 2. Illustrate the CRF algorithm for a specific example in Aspect Extraction task.

As shown above, prediction for output Y_6 with observation X_6 will be based on the set of observations in X and, importantly, the value of Y_5 that has been predicted before. Also intuitively, we can see Y_4, Y_5 are adjectives, Y_6 is a noun so Y_6 will get a high probability of being an aspect, that's why CRF is the perfect solution for Aspect Extraction.

Pipeline

Based on contributions of the AI research team about cross-sentence, our work will create a pipeline, which consists of 2 phases: training and testing.

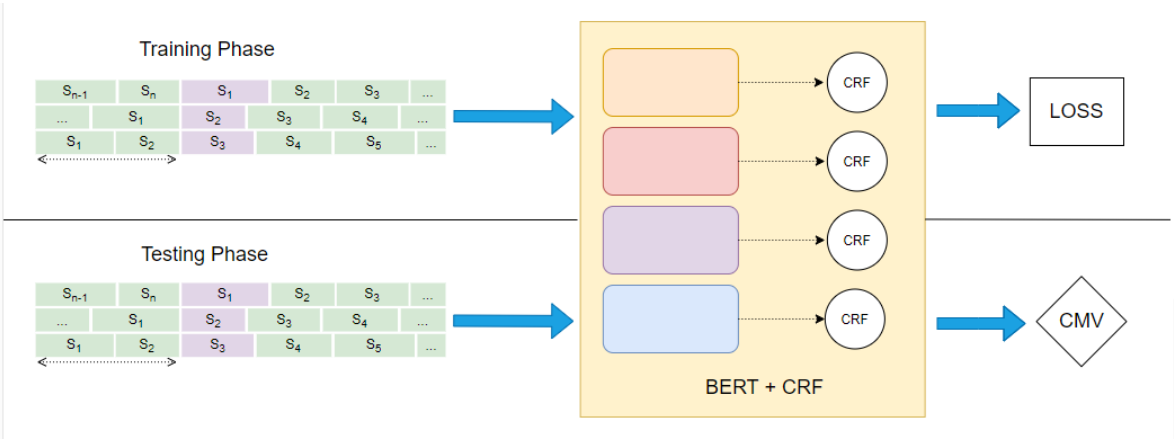


Figure 3. Illustrate the proposed pipeline, including 2 phases of training and testing with the difference in output section.

For each training, we place the sentence of interest at a certain position in the BERT window with a fixed size of 100, then add other sentences in the same data set sequentially at the left and right. Note that the labels of the training data will also be processed according to the above mechanism. (See Figure 3).

At prediction stage, sentences will also be processed in cross-sentence mechanism, except their labels remain the same. The difference is that prediction results will then be passed through a predictive model called the CMV by the research team. Accordingly, prediction results for sentence of interest will be calculated in two ways, the first way is based on the frequency of the label which predicted on each word in sentence of interest and itself but appearing in other predicted sentences that it is not the sentence of interest. And the second way is to calculate the sum of the probability of predicting each word in the sentence of interest and itself but appearing in other predictive sentences and select maximize value representing for the label. (See Figure 3)

6. EXPERIMENTS AND RESULTS

Table 1. SemEval Dataset

	Train		Test	
Dataset	Sentences	Aspects	Sentences	Aspects
Laptop 2014 [21]	3045	2358	800	654
Restaurant 2016 [22]	2000	1743	676	622

To execute the task, we relied on codebase [8] and modified it to accommodate library updates when building the multiple BERT aggregation model. Besides, we also use data process functions of the author group [9], build a cross-sentences dataset with some changes in the naming of variables and functions. For all the remaining model parameters when training, our team keeps the same to ensure the evaluation results show objectively. The model is trained with the last 4 layers of BERT, using Adam optimizer, learning-rate 3e-5, and batch size 16. One slight difference is that we train on GPU (GeForce) GTX 1660S with 6 GB of memory instead of a more powerful GPU because of resource constraints.

Dataset: SemEval is set of a quite famous datasets in Aspect Extraction as well as Aspect -based Sentiment Analysis problem. To perform the Aspect Extraction operation, two datasets are used, SemEval 2014 Laptop and SemEval 2016 Restaurant. Each dataset includes reviews, comments on the subject as well as labeled aspects for each review. (See Table 1)

Analysis: When performing the experiment, we train the model with 10 epochs and observe that the error value varies through each epoch. As shown in the figure below, in both laptop and restaurant sets, at the first epochs, the loss values in both training and validation set are very high, starting from the 4th epoch the model gradually enters convergence point with loss value in training set decreases and at 6th epoch, loss value in validation set is also stable. (See Figure 4.a, 4.b)

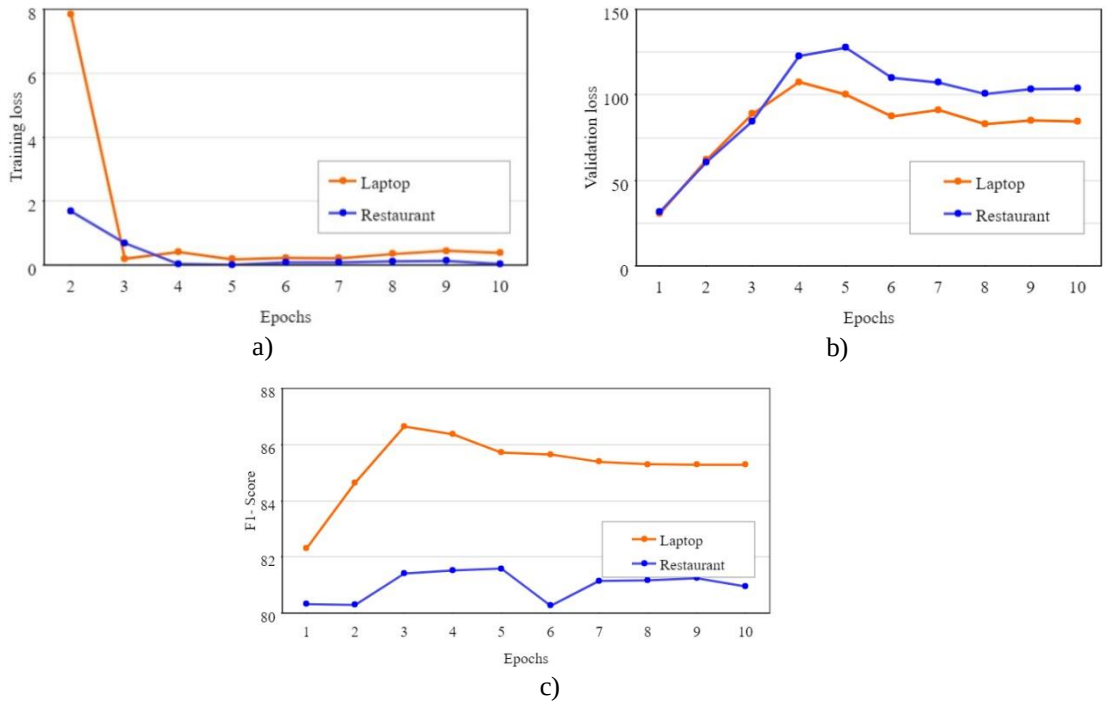


Figure 4. The performance of the pipeline when executing the Aspect Extraction task is measured by the change in the number of epochs during the training period. a) Change on training loss. b) Change on validation loss. c) Change on precision F1

In addition to observing loss value, we also observe the accuracy of prediction on test set when training with different epoch. It can be seen that the accuracy in both the laptop and restaurant will reach the best value when the model is trained from 4 to 6 epochs and will deteriorate if epoch is too high, meaning that it has passed the convergence point. (See Figure 4.c).

Table 2. The table compares evaluation results according to the accuracy F1. The number in bold represents the highest value. Scores not in bold are quoted from [8]. Scores in bold are the superior results of our recommended pipeline. Each result in the table is average of 9 runs.

	Laptop - F1	Restaurant - F1
BERT	79.28	74.10
BERT-PT (30 epochs) [23]	85.93	82.64
P-SUM (4 epochs) [8]	85.94	81.99
Our P-SUM + CMV (4 epochs)	**86.17**	**82.66**
Our P-SUM + CMVP (4 epochs)	**86.28**	**82.71**
Our P-SUM + offset 0 (4 epochs)	**86.17**	**82.21**
Our P-SUM + offset 32 (4 epochs)	**86.29**	**82.80**
Our P-SUM + offset 64 (4 epochs)	**85.96**	**82.67**
H-SUM (4 epochs) [8]	86.09	82.34
Our H-SUM + CMV (4 epochs)	**86.41**	**82.61**
Our H-SUM + CMVP (4 epochs)	**86.43**	**82.83**
Our H-SUM + offset 0 (4 epochs)	**86.24**	82.12
Our H-SUM + offset 32 (4 epochs)	**86.19**	**83.06**
Our H-SUM + offset 64 (4 epochs)	**86.29**	**82.96**

Result: We have tested a lot of measurements to compare with the original model [8]. As be exposed from table 2, with the same number of epochs of 4, most of the results from our pipeline are superior to previously available models. Here, training dataset was processed by cross-sentence mechanism with an offset of 0, while the test dataset is processed and predicted according to CMV (Contextual Majority Voting) and CMVP (Contextual) models. Majority Voting Probability) [9] described in the previous section. Offsets 0, 32, 64 in table 2 are positions of the BERT window where a sentence of interest begins when processing test data. In particular, when performing with P-SUM + offset 32 architecture, accuracy of F1 on our restaurant dataset (82.80) is 0.81 better than P-SUM model (81.99).

7. CONCLUSION

We demonstrated a pipeline with a combination of BERT customization using hidden multilayer integration with a solution that augments the data during training and prediction. We also present a very diverse set of evaluation results based on how the samples in the data are combined. Our idea is very simple, but the obtained results have surpassed the results of previous studies based on the advanced BERT model when executing Aspect Extraction task.

REFERENCES

[1] https://nlp.stanford.edu/software/ tagger.shtml

[2] Wu, Yuanbin, Qi Zhang, Xuan-Jing Huang, and Lide Wu. "Phrase dependency parsing for opinion mining." In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 1533-1541. 2009.

[3] Ghahramani, Zoubin. "An introduction to hidden Markov models and Bayesian networks." In *Hidden Markov models: applications in computer vision*, pp. 9-41. 2001.

[4] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[5] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[6] Kantor, Yoav, Yoav Katz, Leshem Choshen, Edo Cohen-Karlik, Naftali Liberman, Assaf Toledo, Amir Menczel, and Noam Slonim. "Learning to combine grammatical error corrections." *arXiv preprint arXiv:1906.03897* (2019).

[7] Kanerva, Jenna, Filip Ginter, and Sampo Pyysalo. "Dependency parsing of biomedical text with BERT." *BMC bioinformatics* 21, no. 23 (2020): 1-12.

[8] Karimi, Akbar, Leonardo Rossi, and Andrea Prati. "Improving BERT Performance for Aspect-Based Sentiment Analysis." *arXiv preprint arXiv:2010.11731* (2020).

[9] Luoma, Jouni, and Sampo Pyysalo. "Exploring cross-sentence contexts for named entity recognition with BERT." *arXiv preprint arXiv:2006.01563* (2020).

[10] Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. "Opinion word expansion and target extraction through double propagation." *Computational linguistics* 37, no. 1 (2011): 9-27.

[11] Smeaton, Alan F., Fergus Kelledy, and Ruairi O'Donnell. "TREC-4 experiments at Dublin City University: Thresholding posting lists, query expansion with WordNet and POS tagging of Spanish." *Harman [6]* (1995): 373-389.

[12] Shimada, Kazutaka, Ryosuke Tadano, and Tsutomu Endo. "Multi-aspects review summarization with objective information." *Procedia-Social and Behavioral Sciences* 27 (2011): 140-149.

[13] Wang, Wenya, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. "Coupled multi-layer attentions for co-extraction of aspect and opinion terms." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1. 2017.

[14] Huang, Binxuan, and Kathleen M. Carley. "Parameterized convolutional neural networks for aspect level sentiment classification." *arXiv preprint arXiv:1909.06276* (2019).

[15] Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. "Double embeddings and cnn-based sequence labeling for aspect extraction." *arXiv preprint arXiv:1805.04601* (2018).

[16] Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. "Automated Concatenation of Embeddings for Structured Prediction." *arXiv preprint arXiv:2010.05006* (2020).

[17] Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. "How to fine-tune bert for text classification?." In *China National Conference on Chinese Computational Linguistics*, pp. 194-206. Springer, Cham, 2019.

[18] Liu, Yang. "Fine-tune BERT for extractive summarization." *arXiv preprint arXiv:1903.10318* (2019).

[19] Yang, Wei, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. "Data augmentation for bert fine-tuning in open-domain question answering." *arXiv preprint arXiv:1904.06652* (2019).

[20] Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. "What does BERT learn about the structure of language?." In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. 2019.

[21] Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 27–35, 01 2014.

[22] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, et al., "Semeval-2016 task 5: Aspect based sentiment analysis*," in Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.

[23] Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. "BERT post-training for review reading comprehension and aspect-based sentiment analysis." *arXiv preprint arXiv:1904.02232* (2019).

AUTHORS

Anh Khôi Lê - currently studying for a master's degree in information systems at the University of Natural Sciences in Ho Chi Minh City. My current research direction is sentiment analysis and related problems. Social behavior analysis is one of my research fields in near future as well.

PREPARING ANNOTATED DATA ON COVID -19 BY EMPLOYING NAÏVE BAYES

Dipankar Das, Akash Ghosh, AdityaR Rayala, Dibyajyoti Dhar,
Vidit Sarkar,Avishek Garain, Sourav Kumar

Department of Computer Science & Engineering, Jadavpur University, India

ABSTRACT

The on-going pandemic has opened the pandora's box of the plethora of hidden problems which the society has been hiding for years. But the positive side to the present scenario is the opening up of opportunities to solve these problems on the global stage. One such area which was being flooded with all kinds of different emotions, and reaction from the people all over the world, is twitter, which is a micro blogging platform. Coronavirus related hash tags have been trending all over for many days unlike any other event in the past. Our experiment mainly deals with the collection, tagging and classification of these tweets based on the different keywords that they may belong to, using the Naive Bayes algorithm atthe core.

KEYWORDS

Covid-19, Naïve Bayes, Clustering.

1. INTRODUCTION

COVID-19 or SARS-CoV-2 is a severe acute respiratory syndrome (SARS) that emerged from Wuhan, China in December, 2019. It has since resulted in a global pandemic situation and has resulted in over 1.5 million deaths worldwide along with mass disruption of lives for all people around the world. With a pandemic of this scale, social media has had lots to say about the virus ever since its inception. Twitter has seen an unprecedented rise in the number of tweets ever since the pandemic started [1-3] and people from various walks of life have commented on various aspects concerning the virus. These include advice from doctors regarding how to cope with the deadly disease, how people are coping with the new norm of work-from-home, tweets regarding the lockdown and other such. With such a large number of tweets, users are often left searching for what to read as all tweets might not be of interest to them. In this context, segregating the similar tweets together so that the user can read about a particular kind only is essential.

Therefore, the objective of the present work is to classify covid-19 related Twitter data into a specified number of classes and deliver an annotated dataset for the upcoming research communities. In order to accomplish our research goals, we have implemented several varieties of Multinomial Naïve Bayes algorithm from scratch using Python to classify the data into some pre-defined classes on a dataset. We have considered three different strategies to prepare the data as well as to implement the Naïve Bayes models.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.131-147, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111211

In the first case, we have collected 10,000 unique tweets from the whole dataset[1] that ranges from December 2019 till May 2020. The keywords that have been used to crawl the tweets are 'corona', 'covid', 'sarscov2', 'covid19', 'coronavirus'. For this purpose we used Tweepy and Twitter API endpoint.

In the second attempt, we also collected 10,000 tweets. On the initial unlabeled Twitter data, spectral clustering is used to automatically generate class label. This labeled data is used as training data for the classifier and accuracy of the classifier has been calculated.

In the third case also, we have proposed a Naive Bayes [4] based model wherein tweets are classified into various categories or clusters in order to help the user read only a particular type of tweets. In order to evaluate the performance of the system, a dataset which is an in house dataset containing 1000 tweets exclusively related to Covid19 was used. We have collected a dataset of roughly 1000 tweets by parsing Twitter data and collecting the relevant tweets. These tweets are then allotted into 10 classes by employing the spectral clustering method [5] on the cosine similarity of the tweets among themselves. Here, the number of classes is finalized through exhaustive experimentation as described in Section 4.

Not only to prepare a gold standard dataset, we have also proposed the Naive Bayes method which is indeed a commonly used simple but effective classifier for phrase classification. It is based on the commonly known Bayes' Theorem [6] in probability wherein in this case, the probability of each of the tweets belonging to a certain class is obtained. Thus, in a nutshell, the chief highlights of our work include:

- Proposed three novel in house datasets of 21000 tweets related to COVID-19
- Implemented three modified versions of Naive Bayes algorithm from scratch to classify the test set into appropriate classes.
- Performed exhaustive experimentation to finalize our methods which include choosing optimal number of clusters, choosing optimal clustering method, and tuning our Naive Bayes algorithm(s) to fit our purpose.
- Obtained competitive accuracies upon testing, thus proving the robustness of our dataset as well as our method.

The rest of the paper is organized as follows. Three dataset preparation strategies are discussed in Section 2, 3 and 4 respectively whereas various implementations of Naïve Bayes are described in Section 5. Section 6 illustrates the implications of different experiments along with results. Finally, Section 7 concludes the paper by mentioning future tasks.

2. DATASET 1

2.1. Background

Data Pre-processing: The steps that were used on the tweets are as 1) Removal of user, 2) Removal of URLs, 3) Removal of punctuations, 4) contracting unnecessary white spaces, 5) Replacing emoticons with corresponding meaning, 6) Partitioning hash tags etc.

Manual Reviewing: Conducting some manual reviewing we found the following observations; 1) Tweets with four or less words are likely linked to some news article and most of them didn't really convey the whole idea. So we discarded those tweets. 2) Some tweets which were replies to

[1]https://ieee-dataport.org/open-access/english-language-tweets-dataset-covid-19

other tweets sometimes didn't make sense. 3) There were some tweets which were paged like (3/3). Tweet (2/3) was after a lot of tweets. 4) Some tweets belonged to multiple classes. For example: A tweet that contains the statsof recoveries and death could be classified as both news and health.

Data Embedding: To convert the sentences into embedded vectors, TF-IDF was used, with 9774 documents.

Label Identification: A map with key as word in a tweet and value as the count of occurrences in the dataset contained nearly 30000 labels. The map was sorted in descending order, and a list was prepared, using these words. We aim to classify tweets under various classes generated during the "Covid-19 pandemic" in the months of July to August, 2020. Hence, words like "covid19","coronavirus","corona" have been filtered out as they had the highest number of occurrences. From the prepared list, 18 words which have occurred in at least 200 tweets were selected to be the final labels list. The list hence prepared consists of only unigrams.

2.2. Approach

We have chosen three different approaches to tag the given data and manually verify:

Clustering using K-RMS [7]: At first a map is generated between PCA applied embedded data and data cluster points. By that a text file is generated where one can see which tweet is in which cluster. After that labelling of 18 clusters, the output is generated label of 6000 tweets.

The algorithm, "K-RMS" is devised such that it solves issues like the handling signed data problem. It also decreases the number of iterations and increases the accuracy to a great extent. It is observed that if RMS (Root Mean Square) value is used instead of average value, it is expected that the number of iterations will decrease significantly for large datasets. This is because RMS value is much more exact and fast converging in every field of science be it chemistry (VRMS or Root Mean Square Velocity) or some other fields like electrical circuits, etc. It also takes care of negative values in datasets. The degree of changes that takes place during the workflow of the algorithm is lesser compared to that when average value is used.

Cosine Similarity: The vectors for all the tweets and the 18 labels were obtained using TF-IDF technique. Depending on the cosine-similarity value of each tweet's vector and label, if non zero then the tweet was classified under the label giving non-zero value. With this approach labels which are directly represent in the respective tweets is given more preference.

On fitting the processed tweets we obtained vocabulary whose size was 24650 and based on the 18 labels that we had taken into consideration the cosine similarity was done on the vectors of dimensions 1x24650. For each tweet, out of 18 labels the one whose cosine similarity was the highest was considered to be the label that needs to be assigned for the tweet. This strategy was used to label all of the data. The training data was 60% of all labelled tweets and rest formed the test data. We had the idea in mind that if tweet X can label "news" and has the word "news". Another tweet Y which does not have the occurrence of word news but also needs to be labelled "news" by the doing cosine similarity we will likely have the same structure for the sentence (by structure we mean words used and basically the way a tweet is presented for a particular class/label).

Multi Class Annotation: As a tweet might have more than one label, tweets have been labelled as: -> 'label(1) label(2) label(3) … label(N)' depending on whetherthe tweet contains the label(i). Using this approach, 217 odd classes with multiple labels, hereafter referred to as

S-Classes, have been obtained. Many tweets didn't have any of the labels in them. Thus, those tweets have been removed from the training dataset. The number of tweets in the labelled dataset is found to be 3646.

The tweets are transformed into vectors using TF-IDF technique. Multinomial Naive Bayes and Complement Naive Bayes models are used to fit the dataset and prepare it for testing. For testing the model, we considered two directions: 1) Exact Prediction and 2) Subset Prediction. For exact prediction, the predicted class and labelled class are compared on equality. For subset prediction, it is determined if the predicted class is a subset of the labelled class on the basis of (Predicted Class - Labelled Class) ['-' being set difference operator here]. If the operation will return a null, it is considered as a correct prediction. The data was trained using 5 - fold cross validation and the best model was picked to obtain the results. We also present the listof labels which were most incorrectly predicted.

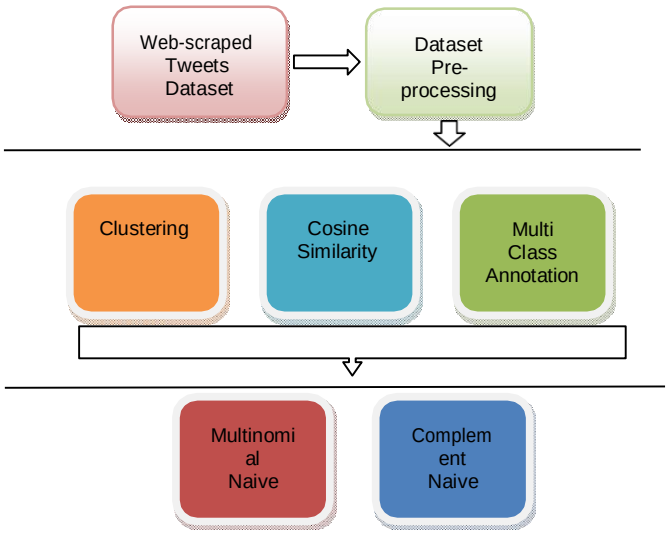


Figure 1. Workflow of Preparing Dataset 1

3. DATASET 2

We have collected 1000 relevant clean tweets related to COVID-19. Our intention was to label the tweets in order to fit them for supervised learning. For this, we have used spectral clustering as well as k-means clustering on the basis of the cosine-similarity matrix generated on the tweet dataset.

In order to obtain the cosine-similarity matrix, we need the word count of the words in each document. We first generate a sparse matrix M of dimension $m \times n$ where m is the total number of tweets and n is the total number of unique words. M_{ij} is an integer value denoting the frequency of the j_{th} word in the i_{th} tweet. For the sparse matrix M, each row indicates an n-length vector. For example, let us consider the r_{th} rowbe $[M_{r0}, M_{r1}, M_{r2}, \ldots, M_{r(n-1)}]$ and let us consider the s_{th} row be $[M_{s0}, M_{s1}, M_{s2}, \ldots, M_{s(n-1)}]$. Now, in n-dimensional space, the two vectors can be considered as two points. The absolute value of the cosine of the angle between the two vectors in the n-dimensional space is considered as the cosine similarity value of the two particular tweets.

Through this process, we get an $n \times n$ size cosine similarity matrix S, where S_{ij} denotes the similarity between the i_{th} tweet and the j_{th} tweet $(0 <= S_{ij} <= 1)$. The larger the value of S_{ij}, the more similar the i_{th} and j_{th} tweet are to each other. Clearly, S is a symmetric matrix. On S, we have done two types of clustering for experimentation and ultimately choose the better one. The types of clustering experimented with are described as follows:

K-means clustering: Here we have just applied the *k*-means clustering algorithm, where the distance between the two *ith* and *jth* data points are considered as the value of S_{ij} or S_{ji}, as both are equal.

Spectral clustering: Spectral clustering is an unsupervised clustering algorithm. It treats the data points as nodes of a graph. The similarity between the data points are calculated using some metric. The technique makes use of the eigenvalues of the similarity matrix and then performs dimensionality reduction. The clustering is formed in fewer dimensions [8]. The steps are as follows.

a) Here the algorithm generates an undirected graph of n nodes, considering S_{ij} as the weight of the edge between i_{th} and j_{th} node.
b) From S, we generate D as a diagonal matrix, where D_{ii} is the sum of all S_{ik} for all k between 0 and $n-1$.
c) Now we apply the formula to obtain the Laplacian Matrix L where $L = D - S$.
d) From L we calculate normalized Laplacian matrix $L_{norm} = D^{(-½)}LD^{(-½)}$
e) From L_{norm} we calculate first k eigenvectors $v_1, v_2,..., v_k$.
f) Let U be the matrix containing the vectors $v_1, v_2,..., v_k$ as columns.
g) For $i = 1,..., n$, we take the i_{th} row of U as its feature vector afternormalizing to norm 1.
h) Then, we cluster the points with k-means into k clusters $C_1,..., C_k$

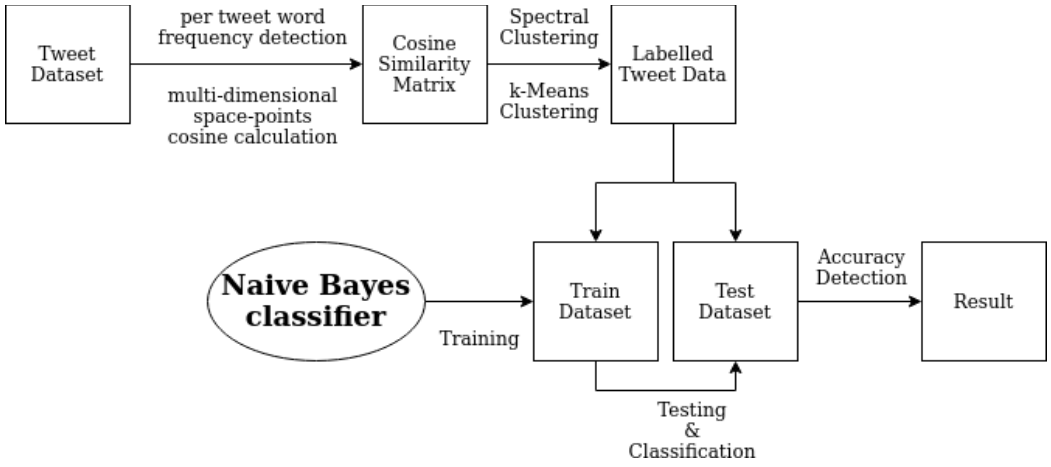


Figure 2. Flowchart representing the module-wise workflow of Dataset 2

4. DATASET 3

The dataset consists of 10,000 tweets. The data crawling is done using Tweepy, which is an easy-to-use Python library for accessing the Twitter API. For accessing Tweepy, authentication to Twitter API is required. The data is then cleaned and preprocessed by removing urls, emojis and special characters and then stored in a .csv file. This code collects the 10,000 most recent tweets with respect to the search words, and filters out retweets. Then the preprocessing is done and the

data is stored in the .csv file.

Then the first 100 tweets are manually labelled and stored in a file covid- labeled-data.csv. Then Naïve Bayes is applied to the data to generate the other labels. This is stored in a file covid-test-data.csv. Spectral clustering has been done on the datato test accuracy of the algorithm. The input for spectral clustering is covid-cleaned- data.csv. The labels are auto-generated using this method and are used to verify with the Naïve Bayes algorithm.

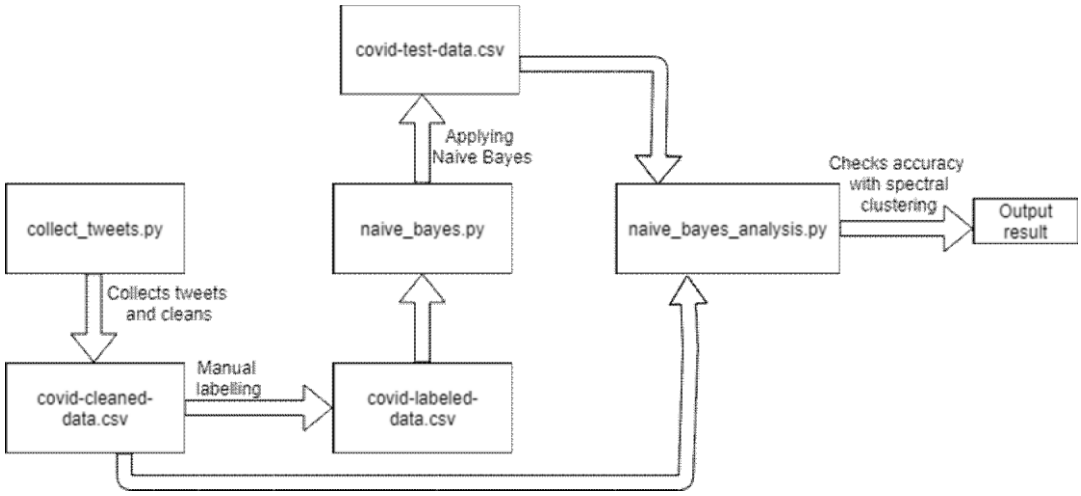


Figure 3. The structure diagram showing module-wise workflow of Dataset 3

5. NAÏVE BAYES MODEL

Multinomial Naïve Bayes is a classification technique based on Bayes' Theorem. In this classifier, "bag of words" document representation is used. The words and phrases are the features. The Naïve Bayes model assumes that the feature probabilities are independent given a class c. The conditional probability of belonging to a class c given the document d is calculated. The Naïve Bayes algorithm is given in detail in [1].

Bayes Theorem states that, for two events A and B, if probability of event A occurring is P(A), probability of event B occurring is P(B), and probability of event A occurring given that event B has already occurred is P(A/B) , then probability of eventB occurring given that event A has already occurred is P(B/A) = (P(A/B)*P(B))/P(A).

$$P(A \cap B) = P(A/B).P(B) = P(B/A).P(A)$$
$$P(B/A) = (P(A/B).P(B))/P(A)$$

, where $P(A \cap B)$ is the probability of occurring A and B together. Now, in our context, suppose c is a class and d is a document. Now given the document d, the probability of it belonging to class c is:

$$P(c/d) = (P(d/c).P(c))/P(d)$$

We will have multiple classes $\{c_1, c_2, c_3, \ldots c_N\}$ and we have to figure out inwhich class out of these, our given document d belongs to. So, if out of N classes, the probability corresponding to the i^{th} class $P(c_i/d)$ is highest, then we say that document d belongs to the ith class ci. One thing to notice here is that, while calculating the probability for each of the classes, the term P(d) in the denominator on the right side is exactly the same. So, we can simply ignore that term, as it won't

relatively affect the results. So the our equation becomes: $P(c/d) = P(d/c).P(c)$.

Now, our document d may contain M words x_1, x_2, x_3… x_M. So our equation becomes:

$$P(c/d) = P(d/c).P(c) = P(\{x_1, x_2, x_3...x_M\}/c).P(c) = P(x_1/c).P(x_2/c)...P(x_M/c).\ P(c)$$

We call this algorithm Naive Bayes because we simply neglect any interrelation between the words in the document. So, we calculate the probability of each class P(c) from the given training dataset. Then, we select the "Bag Of Words" which, in our case, are all the words that appear in the training dataset. Next, we calculate the probability of each unique word from the bag of words belonging to a particular class c, i.e, P(x/c). Finally, we apply the Bayes theorem on the testing data document and find the class with highest probability and then we check how accurately we are able to predict.

$$C_{map} = \text{argmax}_{c \in C} [\ P(c).\Pi_{x \in d} P(x/c)]$$

The Naïve Bayes experiments have been conducted on these three varieties of datasets and important observations were grouped into three different sections described as follows.

6. EXPERIMENTS AND RESULTS

6.1. Observations 1

The general observation is that in all the different models, complement Naive Bayes works better relatively, because the model is designed to handle datasets which are class-imbalanced, and can be visualized from the above given histogram. The accuracy obtained using the cosine similarity label tagging is higher compared to the remaining two methods, since the classifier guesses the label based on its existence in the tweet. Also, in most cases this was the expected tag of the corresponding tweet. Except in some abstract cases where the label of the tweet isn't occurring in the tweet itself.

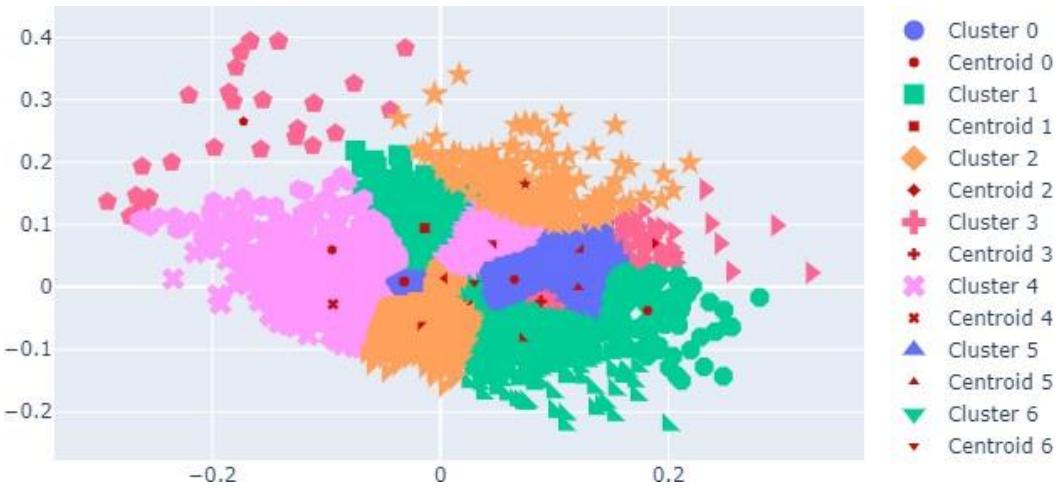


Figure 4. K-RMS Clustering results with 18 clusters on Dataset 1

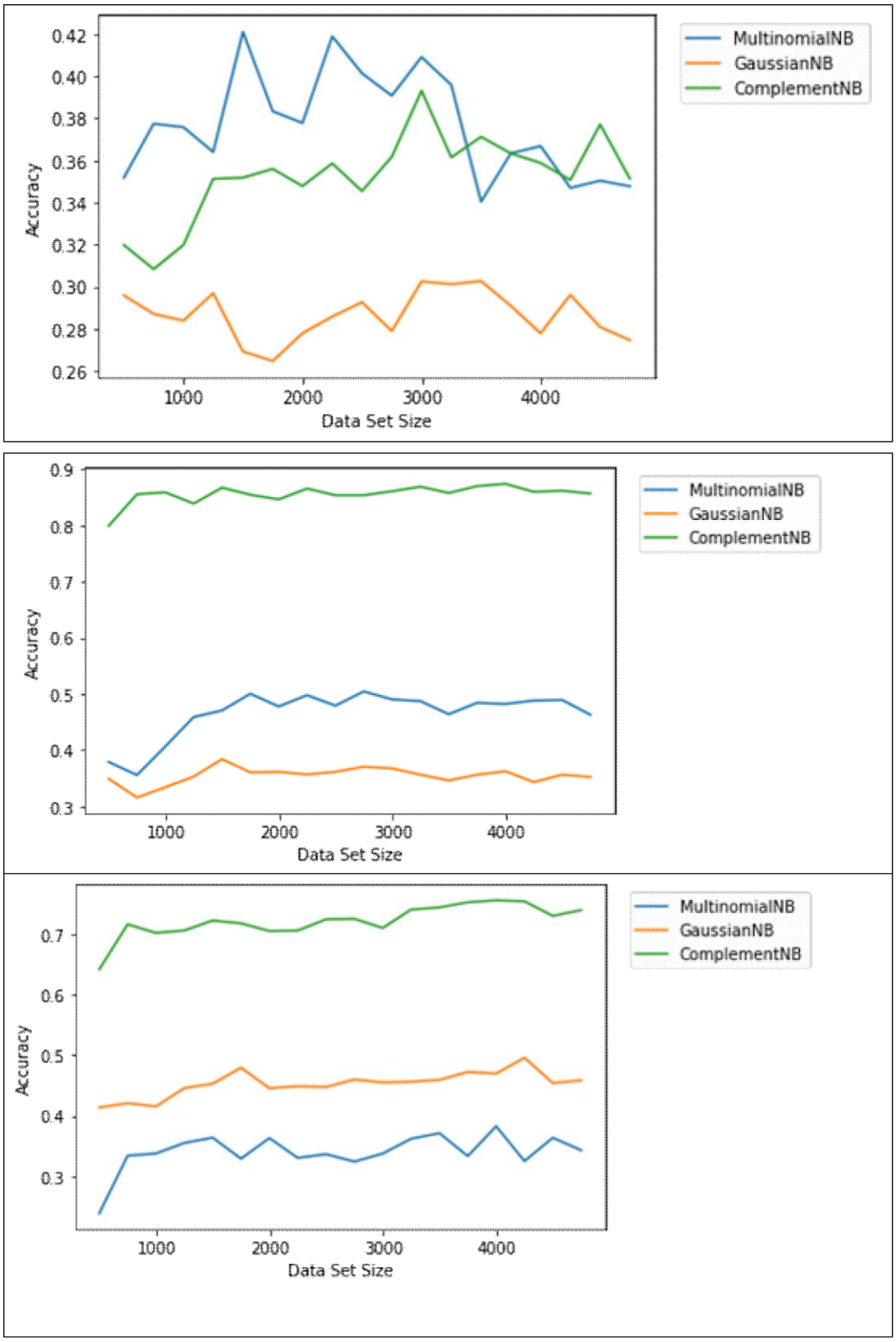


Figure 5. Performances of Multinomial, Gaussian and Complement Naïve Bayes onDataset 1

The highest accuracy obtained using 5 - fold cross validation is 88 per cent, labelled using Cosine Similarity and model used is Complement Naive Bayes (as shown in Figure 4 and Figure 5). The clustering model wasn't very accurate with the cluster formation since the classifier seems to be confused while assigning a specific label. The different methods have predicted some words most incorrectly, for example, using the clustering method, economy is the most incorrectly labelled word. It is due to the fact that the data belonging to the economy class is not properly clustered under one label, and belongs to multiple clusters. Also, mask is the most incorrectly labelled word for cosine similarity; it is due to the fact that it doesn't generally occur within a tweet. The plots comparing the different methods of tagging, are illustrated below, for comparison of performances, over increase in the size of data.

Most Incorrectly Predicted Words [Complement NB]	
KNN Cluster	**Cosine Similarity**
Economy	Mask
Outbreak	Death
Mask	China

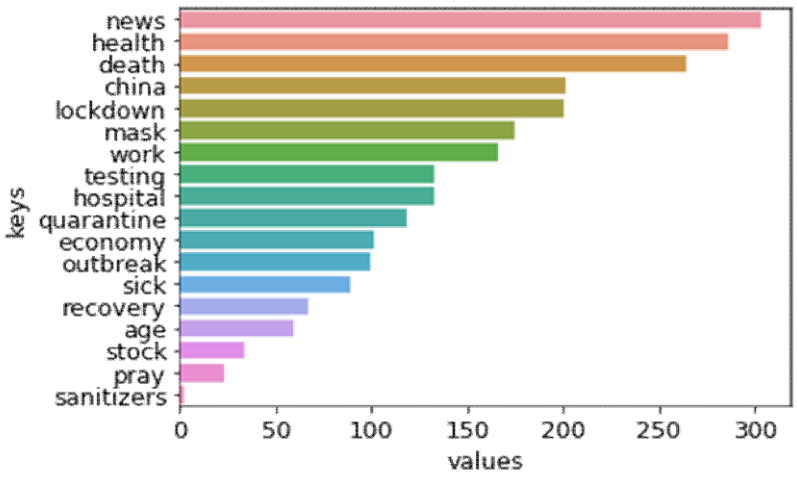


Figure 6. Predicted Results with frequencies on Dataset 1

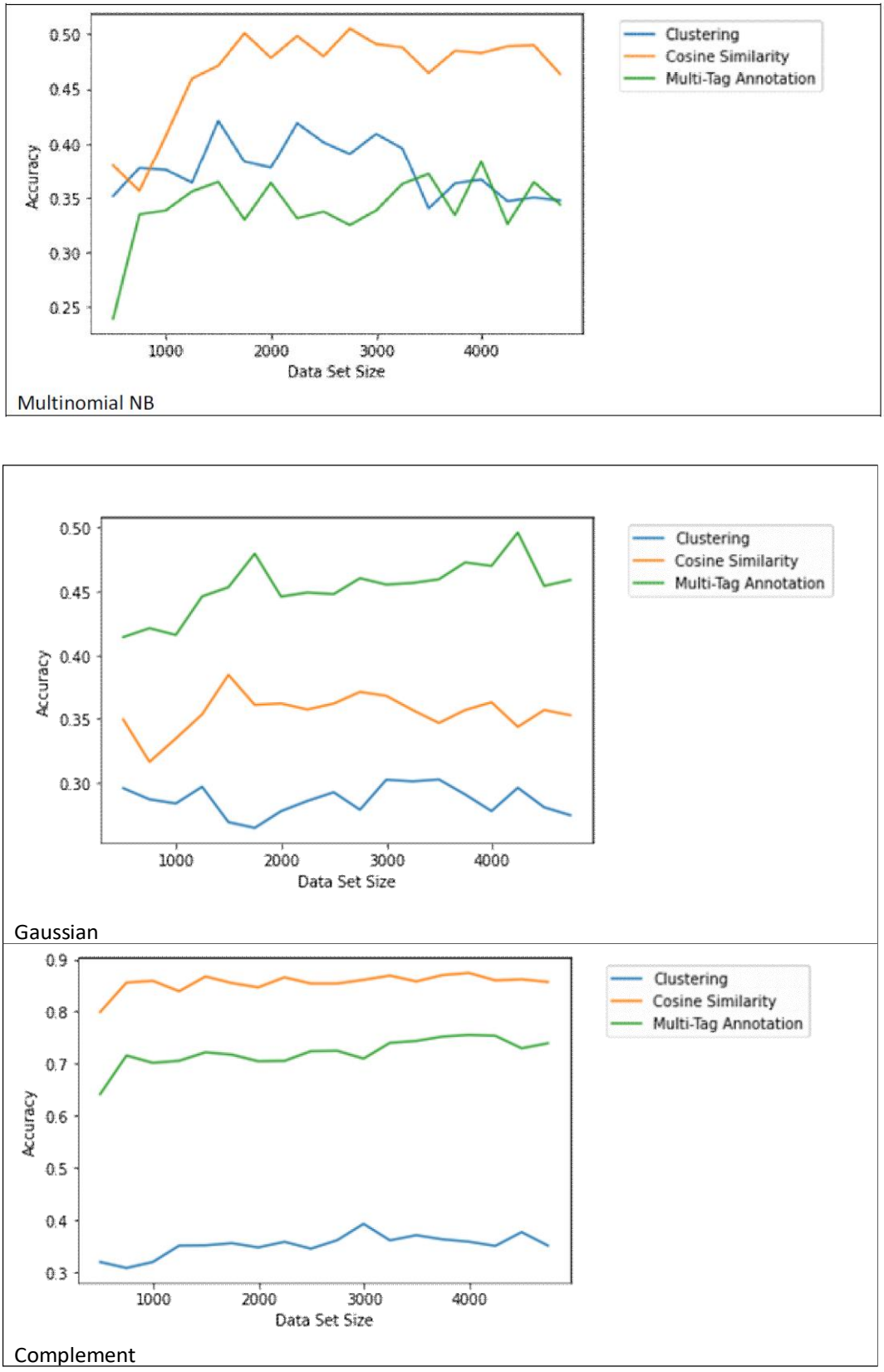


Figure 7: Performances of Multinomial, Gaussian and Complement Naïve Bayes withrespect to three techniques (clustering, cosine and multi-tag) on Dataset 1

6.2. Observations 2

In this experiment,

- For each kind of clustering we have prepared datasets consisting of 5 to 15classes.
- For each such class, 70% of the total dataset has been considered for trainingand the rest for testing.
- The results of all these experiments are represented in a comprehensive tabularform in Table 1.

Table 1: Accuracies obtained for all 5-15 classes for both Spectral and K-meansclustering on Dataset 2

Number of classes	Accuracy (%)	
	Spectral clustering	K-means clustering
5	66.33	64
6	65	66.67
7	69	66
8	68	63.33
9	69.33	67.33
10	73.67	64
11	69.33	65.33
12	69	66.67
13	72.33	60.67
14	70.33	54
15	62	69.33

To further validate our method, we have tested the entire process on another publicly available dataset. This dataset [7] consists of a collection of 100 tweets which is divided into train and test sets in the ratio of 1:4. Since the dataset is already available, we know the most optimum number of classes that we should get as output (4 in this case). Hence, an answer near 4 will indicate the robustness of our method. As before, we test across a variety of classes (2-8) using the two clustering methods andtake the best accuracy as our answer. The results are illustrated in Table 2.

Table 2: Accuracies obtained for all 2-8 classes for both Spectral and K-meansclustering on the second dataset

Number of classes	Accuracy (%)	
	Spectral clustering	K-means clustering
2	95	95
3	85	100
4	95	95
5	100	95
6	100	90
7	95	95
8	90	90

From Table 1, it is evident that in most of the cases, Naive Bayes achieves more accuracy for Spectral clustering than K-means clustering. Except for 5-class and 15- class datasets, K-means clustering has more accurate results than that of Spectral clustering. The highest accuracy for Spectral clustering is 73.67% when there are a totalof 10 classes whereas a highest accuracy of 69.33% is achieved for the 15 class case of K-means clustering. On average, Naive Bayes achieves more accuracy for Spectral clustering than K-means clustering because of the eigenvector set generation with the help of the normalized Laplacian matrix in spectral clustering algorithm, which actually helps to more effectively cluster the data points in an n-dimensional space, in comparison to simple k-means clustering, where only cosine matrix element values are taken as the distance between two data points. In spectral clustering, k-means clustering algorithm runs at the last stage with respect to the eigenvector set. For datasets containing a wide variety of classes, spectral clustering (k-partitioning of connected graph) is more effective. Since the accuracy for 10 classes with spectral clustering is the highest, we consider that to be the output of our method. The top 5 words of each of the 10 classes are illustrated in Table 3.

Table 3: Top 5 most frequently used words in each class for 10 classes with spectralclustering along with the suggested class labels on Dataset 2

Class no.	Most Frequent Words	Suggested Class Label
1	china, pandemic, cases, india, covid19	covid impact on asian countries (withlarge population)
2	funds, news, order, until, vaccine	covid-vaccination
3	cases, deaths, people, coronavirus, covid19	covid-deadliness
4	stock, trading, warns, york, coronavirus	impact of covid on economy
5	working, nurses, salary, staff, sir	occupation and health-workers related
6	like, pictwittercom, corona, covid, coronavirus	general covid-related information (occasionally pictorial)
7	coronavirus, lot, masks, people, see	masks and covid awareness
8	affected, flood, jumps, spike, tally	covid statistics
9	positive, report, tested, thursday, washington	covid testing in washington
10	they, my, we, you, covid19	Personalised Covid-related tweets

From Table 2, we note that our method is indeed robust as for spectral clustering, the optimum value of 100 is reached with number of classes=5 which is very near the optimum value of 4. We also note that the accuracy for 4 classes is 95 which mean that only one sample is classified incorrectly which is negligible. Further, the best accuracies are obtained around 4 classes which denote that our method is logically sound.

6.3. Observations 3

In this particular experimental set up, the first 100 tweets are manually labelled. These 100 labelled tweets have been considered form the training data used for the Naïve Bayes classifier. The tweets are divided into four classes. They are as follows:

1. Covid-19 prevention mechanisms
2. Statistics related to Covid-19
3. Vaccine or medications related to Covid-19
4. Tweets related to other topics

Using manual labelling, 11 tweets were classified as class-1, 19 as class-2, 14 as class-3 and 56 as class-4. So P(1)=0.11, P(2)=0.19, P(3)=0.14, P(4)=0.56. The words and phrases related to covid-19 are stored in a list. They consist of monograms, bigrams and trigrams and were hard-coded into the list. The following is the list of words and phrases used as the features:

Table 4. Monograms, bigrams and trigrams used on Dataset 3

Monograms	**Bigrams**	**Trigrams**
Case/s, Death/s	Social distancing	Case Fatality Rate
Test/s, Hospitalization/s	Recovery rate	
Mask/s, Pandemic	Fatality rate	
China, Recovery, Lockdown	Contract tracing	
Quarantine, Vaccine/s, Moderna	Herd immunity	
Pfizer, Immunity		

The first 100 tweets were used form the training data and the remaining 9900 tweets form the test data. In the second experiment, spectral clustering is done on the dataset. The first 200 tweets are divided into four clusters using spectral clustering algorithm. The cosine similarity between the tweets is used to create the similarity matrix. The clusters are generated from the cosine similarity matrix. The data is now trained with first 100 tweets and tested for the next 100 tweets. A counter is set so that when the classes obtained by the Naïve Bayes classifier and the spectral clustering match, the counter is incremented. Using the formula for accuracy,

Classification accuracy = Correct predictions / Total predictions

The accuracy is calculated to be 86%.

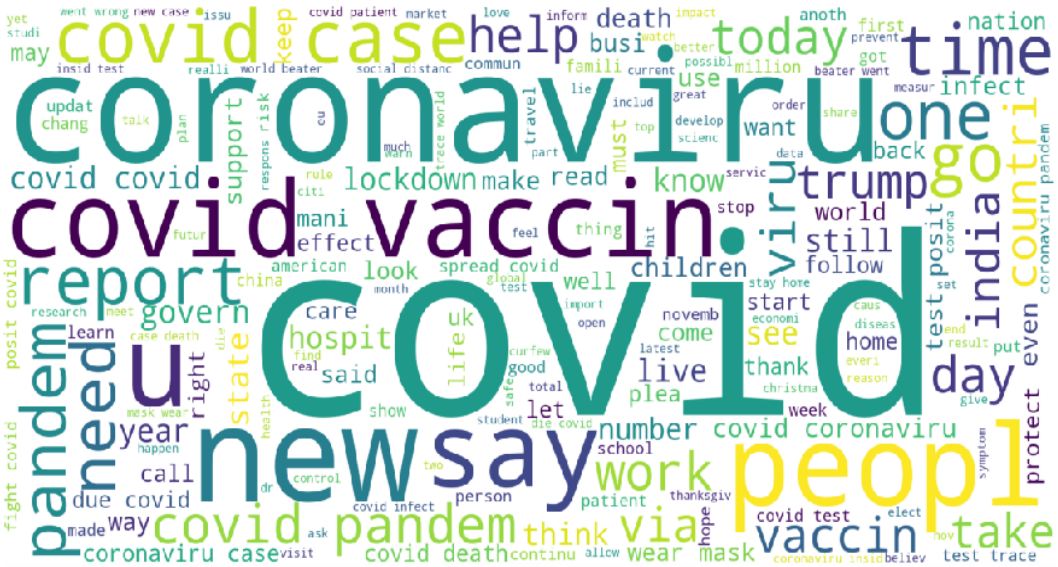


Figure 8. Word Cloud on Covid-19 Dataset 3

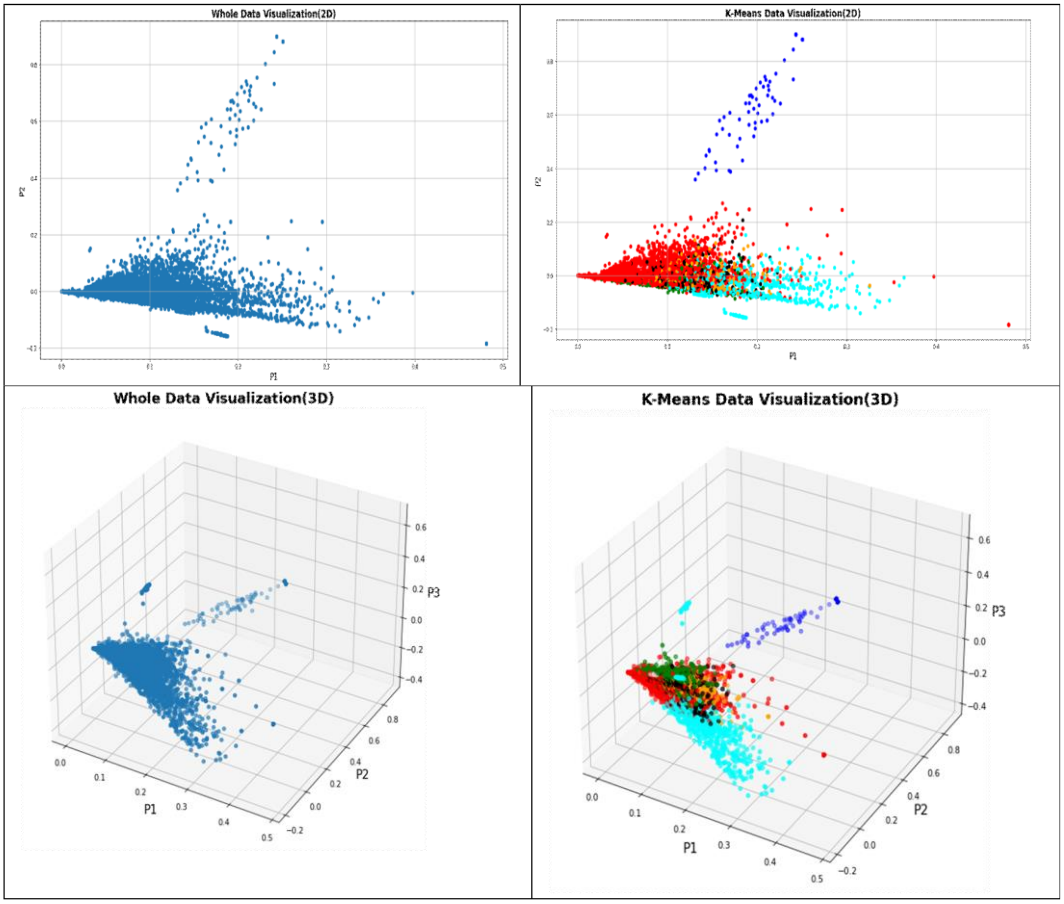


Figure 9. Clustering Results on Dataset 3 in 3D view

K-Means clustering algorithm is used here to label the tweets. The best number of clusters is found using silhouette_score. The number of clusters is varied from 2 to 9. n_clusters = 8 gave silhouette_score = 0.007378162655447713 which is the best. So, number of clusters is taken as 8. Here is different n_clusters vs silhouette_score plot,

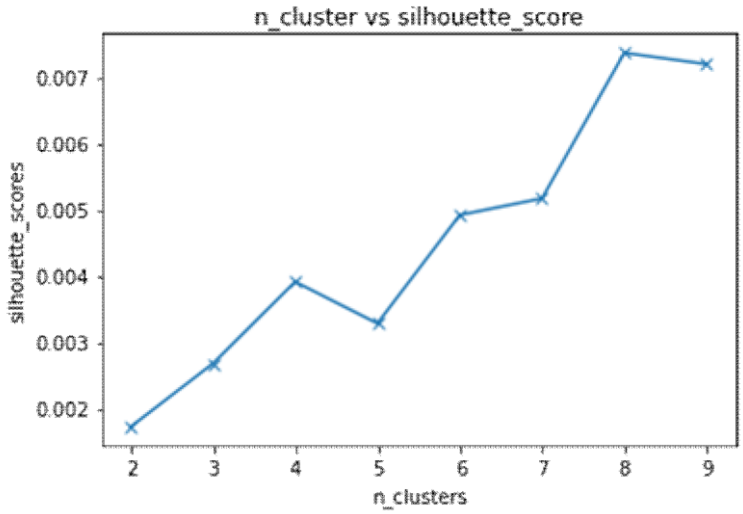
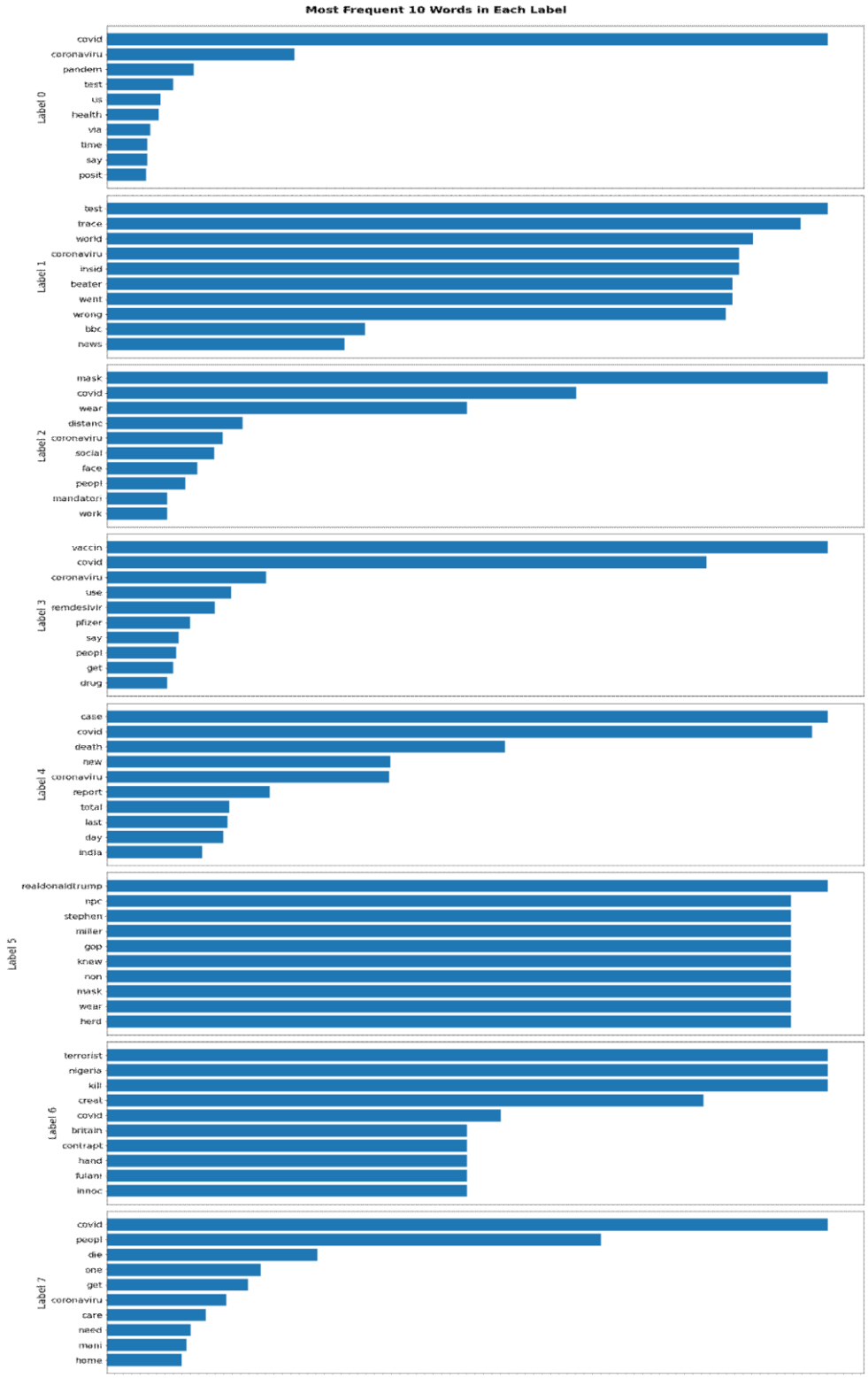


Figure 10. Plot between n_clusters and silhouette_score


Most Frequent 10 Words in Each Label
Label 0
covid
coronaviru
pandem
test
us
health
via
time
say
posit
Label 1
test
trace
world
coronaviru
insid
beater
went
wrong
bbc
news
Label 2
mask
covid
wear
distanc
coronaviru
social
face
peopl
mandatori
work
Label 3
vaccin
covid
coronaviru
use
remdesivir
pfizer
say
peopl
get
drug
Label 4
case
covid
death
new
coronaviru
report
total
last
day
india
Label 5
realdonaldtrump
npc
stephen
miller
gop
knew
non
mask
wear
herd
Label 6
terrorist
nigeria
kill
creat
covid
britain
contrapt
hand
fulani
innoc
Label 7
covid
peopl
die
one
get
coronaviru
care
need
mani
home

7. CONCLUSIONS

In our work, we have developed a Naive Bayes based Algorithm for classification of Covid-19 related tweets. We have first collected an in-house dataset consisting of 1000 tweets by crawling Twitter and collecting tweets related to Covid. Subsequently, we have assigned them different classes using spectral clustering and *k*-means clustering. Then, using Naive Bayes Classifier which we have implemented from scratch, we have classified the tweets into the various classes. In future, we would like to collect more tweets related to Covid so that the classifier can be better trained with a larger dataset to handle the tweets. Further, we would like to test with more classifiers and perform a comparative study with other classifiers in regard to the performance in classifying our dataset.

The multinomial Naïve Bayes algorithm is implemented. We plan to use more clustering techniques like k-means and experiment with different number of classes. Weplan to compare the accuracy obtained by these different methods and find out the optimal number of classes.

ACKNOWLEDGEMENTS

The work is supported by the project "*Sevak- an Intelligent Indian Language Chatbot*" funded by DST-SERB, MeiTY, Govt. of India.

REFERENCES

[1] Sharma, Karishma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. "Covid-19 on social media: Analyzing misinformation in twitter conversations." *arXiv e-prints* (2020): arXiv-2003.

[2] Dimitrov, Dimitar, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. "TweetsCOV19-A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2991-2998. 2020.

[3] Sarker, Abeed, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. "Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource." *Journal of the American Medical Informatics Association* 27, no. 8 (2020): 1310-1315.

[4] Xu, Shuo. "Bayesian Naïve Bayes classifiers to text classification." *Journal of Information Science* 44, no. 1 (2018): 48-59.

[5] Ng, Andrew, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems* 14 (2001): 849- 856.

[6] Joyce, James. "Bayes' theorem." (2003).

[7] https://github.com/gituhin/sentence-classification-naive-bayes-

A SELF-AGGREGATED HIERARCHICAL TOPIC MODEL FOR SHORT TEXTS

Yue Niu and Hongjie Zhang

University of Science and Technology of China, Hefei, Anhui, China

ABSTRACT

With the growth of the internet, short texts such as tweets from Twitter, news titles from the RSS, or comments from Amazon have become very prevalent. Many tasks need to retrieve information hidden from the content of short texts. So ontology learning methods are proposed for retrieving structured information. Topic hierarchy is a typical ontology that consists of concepts and taxonomy relations between concepts. Current hierarchical topic models are not specially designed for short texts. These methods use word co-occurrence to construct concepts and general-special word relations to construct taxonomy topics. But in short texts, word co-occurrence is sparse and lacking general-special word relations. To overcome this two problems and provide an interpretable result, we designed a hierarchical topic model which aggregates short texts into long documents and constructing topics and relations. Because long documents add additional semantic information, our model can avoid the sparsity of word co-occurrence. In experiments, we measured the quality of concepts by topic coherence metric on four real-world short texts corpus. The result showed that our topic hierarchy is more interpretable than other methods.

KEYWORDS

Hierarchical Topic Model, Texts Analysis, Short Texts, Data Mining.

1. INTRODUCTION

Short texts corpus is a kind of prevalent format of texts on the internet, such as titles, comments, microblogs, questions, etc. Applications are interested in discovering knowledge behind the content of short texts. But it has been recognized a challenging task because that the content is unstructured. To analyze the content, ontology learning methods are proposed [1].

Hierarchical topic model is one kind of typical ontology learning model. This model will construct a tree structure in which nodes are concepts and relations are taxonomy relations between concepts. This hierarchy structure can support a wide range of content analysis tasks. The tree structure can provide a global view so researchers can easily understand what are the current research focus and keywords of this focus. Also, the tree structure can support emergency topic detection so customer service agents can quickly know problems of products from the corpus of user feedback. And for users of microblogs or news, the hierarchical topic tree can help them finding contents of their interests by enhancing search engines.

Currently, researchers propose different kinds of hierarchical topic models, such as nCRP [2], nHDP [3], hPAM [4], etc. These models construct tree structures according to the different hypotheses. But for all these models, short texts sets are not easy to deal with. The reason is that

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.149-159, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111212

all these models rely on co-occurrence words and general-special word relations to clustering words into topic. But short texts lack these semantic relations.

Hierarchical model thLDA [5] is proposed to analyze tweets on Twitter. They incorporate word embeddings into a hierarchical model to provide additional semantic information. But all these word embeddings information is generated from Google News, an auxiliary documents corpus. But this auxiliary information may only be suitable for tweets, but unsuitable for other short texts. Word embeddings may be semantic word relations in one domain, but may be non-semantic word relations in other domains. So if the auxiliary corpus is inappropriate, the word embeddings information will bring a lot of non-semantic information into the hierarchical model and will lead to poor performances.

To deal with short texts, another compromised method is proposed at the state of data preparation. This method is called text pooling [6], which draws support from auxiliary information. If short texts have auxiliary information of labels, writers, social relations, etc. This method can incorporate this information by pooling short texts with same labels or writers into long documents. After this data preparation, hierarchical models can deal with these data set as common long documents sets. But this kind of method has some drawbacks. First, the auxiliary information is not very common or easy to acquire. Second, the quality of auxiliary information cannot easily evaluate. Self-aggregated model like PTM [16] generate short texts into long documents to add additional word co-occurrence. But this model cannot provide hierarchical topics.

Inspired by the these methods, our research proposes a self-aggregated hierarchical topic model (shPAM) which aggregates short texts heuristically and does not rely on auxiliary information. We also aggregate short texts into long texts. Long texts can provide a lot of word co-occurrence information and construct additional general-special word relations. So our model can overcome the problems above. We designed long documents as a latent variable. Then we can generate a joint probability distribution with variables: words, topics, long documents, and short texts. Firstly, short texts will implicitly aggregate into latent long documents. Then, we generate topics and topic relations from latent long documents, which are much longer than short texts. In this way, the hierarchical model is generated on latent long documents, avoiding the sparsity of co-occurrence information.

To measure our result, we designed the metric of topic coherence. We adopt PMI score[7] for measuring whether a topic is easy to be interpreted into a comprehensive concept. We compared our model with both baseline models and state-of-the-art models on four different kinds of short texts set. The result showed that the concepts of our model is more coherent compared to other methods.

The remainder of this paper is organized as follows. In section 2, we propose related works. In section 3, we give our model and model inference. We present experimental results in section 4 and conclude our work in section 5.

2. RELATED WORKS

Constructing topic hierarchies is one kind of ontology learning [1]. Ontology learning aims at extracting concepts and relations between concepts from the corpus. Several works only extracting concepts or extracting different kinds of relations. But in our research area, we only extract concepts and taxonomic relations. In this area, methods can be split into two groups. One group methods employ the graphic model which formulates the problem as a joint probability

distribution. Another group method is hierarchical clustering which recursively aggregates layer by layer.

2.1. Topic Model

Different graphic models follow different hypotheses and construct different kinds of hierarchical topic trees. Early proposed methods are hLDA [8] and nCRP [2]. These two methods suggest that one document corresponds to a path in the tree. This suggestion is not very reasonable because one document may correspond to more than one path. Method PAM [9] supposes that only leaf nodes can be translated into concepts. Non-leaf nodes have no explicit meanings. Method hPAM [4] can construct a three-layer topic hierarchy and assume one document may correspond to any topic in the tree. This method can provide only a tree but also a DAG. Method rCRP [10] is very similar to hPAM, but it does not need users to configure the number of topics and the height of the tree. Method nHDP [3] and Ahmed et al[11] have similar models, they suppose that one document corresponds to a sub-tree in the final result. All these methods cannot properly deal with short texts. thLDA [5] is a hierarchical topic model especially for tweet. This model uses word embeddings as the auxiliary information to overcome the sparsity of word co-occurrence. But in this model, if the auxiliary documents are not semantically related with short texts, word embeddings will bring non-semantic information into the model and lead to poor performances.

2.2. Hierarchical Clustering

Hierarchical clustering methods are quite different from the topic models. The graphic model is only a kind of soft-clustering, one document corresponds to several topics with probability. But hierarchical clustering methods cluster documents at every layer. So each document determinately corresponds to one topic at each layer. Bayesian Rose Trees [12] can generate trees with any structure, not only binary trees. Liu et al. [13] construct a topic tree for keywords. This method needs knowledge from auxiliary information. Wang et al.[14] propose a method specialized dealing with content-representative documents such as titles of academic papers. Zhang et al. [15] designed a new combination method specialized for short texts. But their model cannot provide interpretable taxonomic relations as child nodes can be very familiar with father nodes.

2.3. Pooling Method

Pooling methods can help to deal with short texts through auxiliary data. Mehrotra et al. [6] propose a pooling method for tweets by hashtags and can also deal with the corpus that partially has labels. Ahmed et al. [11] also, deal with short texts by pooling them through user id. But auxiliary data are not always available. Besides, pooling through inappropriate auxiliary data may lead to a worse result. PTM [16] is a self-aggregated topic model, but this model cannot generate hierarchical topics.

3. MODEL AND INFERENCE

In this section, we propose our hierarchical model especially for short texts. Firstly, we introduce our model and show how it could supply additional co-occurrence information without auxiliary data. Then, we give the inference method of our model.

3.1. Model

Our model assumed that there are K topics as nodes of the hierarchical tree. The hierarchy has three layers. The first layer is the root topic. The second layer has K_T topics, which are defined as super-topics and root-topic. The third layer has K_t topics as leaf nodes, which are defined as sub-topics and one more topic as super-topic. The total topic number is $K_T + K_t + 1$. For each topic, there is a multinomial distribution over the vocabulary of size V. Also, we assumed that the number of short texts is N_d. The number of words in one short text is represented as N_w. And the number of latent long texts is N_D. There is a multinomial distribution between short tests and latent long texts with parameter φ. So each short text belongs to a long text, and N_D is smaller than N_d. The generation process can be briefly described as follow. Firstly we sample a long text for a special short text. The long text has multinomial distributions over K topics. Then we sample a path of topics in the hierarchy. Having the path, we sample a topic from the path. At last, a word is sampled according to the distribution of the topic.

The generation model is as follows:

1) Sample $\varphi \sim Dir(\alpha)$
2) For each topic z:
   Sample $\eta_z \sim Dir(\gamma)$
3) For each latent long document
   a) For each super-topic:
      Sample $\theta_T \sim Dir(\beta_T)$
   b) For each sub-topic:
      Sample $\theta_t \sim Dir(\beta_t)$
4) For each short document d_i
   a) Sample a latent long document $D \sim Mult(\varphi)$
   b) For each word w_j in short document d_i
      i) Sample a super-topic $z_T \sim Multi(\theta_T^D)$.
         If $z_T = z_{root}$, sample $w_j \sim Multi(\eta_{z_{root}})$
      ii) Else sample a sub-topic $z_t \sim Multi(\theta_t^D)$.
         If $z_t = z_{supe}$, sample $w_j \sim Multi(\eta_{z_T})$
      iii) Else sample $w_j \sim Multi(\eta_{z_t})$

The graphical model can be seen as follows:

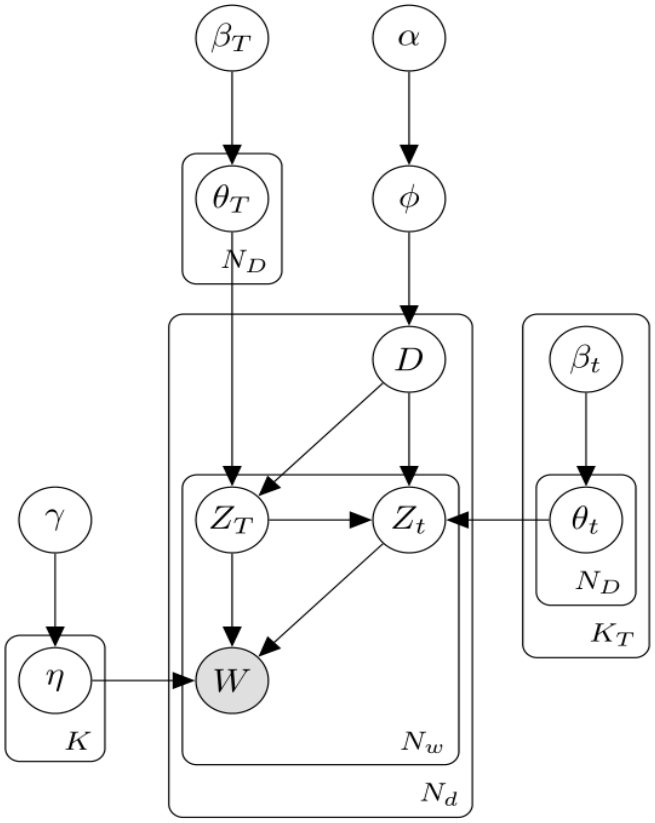


Figure 1. Graphial model of shPAM

The number of latent long documents should be defined carefully. For a short texts corpus, the number of short text N_d is always big. But for each short text, the number of co-occurrence relations M_d between two words are small. Not enough co-occurrence relations will lead to inaccurate results. So in our model, we aggregate short texts by setting $N_D < N_d$. For each long document, the number of co-occurrence relations M_D will be much larger than a single short text, as $M_D > M_d$. Besides, these latent variables are different from latent topics. Latent topics are specific topics with specific concepts. But latent long documents are the aggregations of short texts. A long document can be regarded as a combination of topics. So we should guarantee that $N_D < N_d$ and $N_D > K$. Finally, the graphical model of shPAM can be seen in Figure 1. The solid node W means this variable is all ready known.

In addition, our model proposes a three-layer hierarchical topic model. But the number of layers can be easily extended. We can define sub-sub-topics and recursively processing the topic sampling section in the generation model.

3.2. Inference

As the model is very complex, we train our model by collapsed Gibbs sampling. There are three latent variables for sampling : long document D, super-topic z_T and sub-topic z_t.

Firstly, we sample latent variable D. For every short text m, if it is aggregated into latent long document d, the probability is as follows.

$P(D_i=d|D_{-i},w, z_T, z_t, \alpha, \beta_T, \beta_t) \propto$

$$\frac{n_d+\alpha_d-1}{N+N_d\alpha-1}\prod_{j=1}^{L_m}\frac{1}{n_d+K_{z_T}\beta-J}\prod_{z_T\in m}\prod_{k=1}^{n_m^{z_T}}(n_d^{z_T}+\beta_T-k)\prod_{z_T\in m}\prod_{l=1}^{n_m^{z_T}}\frac{\prod_{z_t\in m}\sum_r^{n_m^{z_T,z_t}}(n_{z_T,d}^{z_t}+\beta_T-r)}{n_d^{z_t}+K\beta_t-l}$$

In this function, we integrate out parameters θ_T, θ_t andφ. n_d is the number of short texts that belong to latent long document d. N is the number of short texts and N_d is the number of latent long documents. L_m is the number of tokens in short text m. n_d is the number of tokens assigned to latent long document d, and K_{z_T} is the number of z_T+1. $n_m^{z_T}$ is the number of tokens belong to short text m with topic z_T, and $n_d^{z_T}$ is the number of tokens belong to latent document d with topic z_T. $n_m^{z_T,z_t}$ is the number of tokens in short text m assigned to z_T and z_t. $n_{z_T,d}^{z_t}$ is the number of tokens of document d assigned to z_T and z_t. $K_{z_t}^{z_T}$ is the number of z_t+1 belongs to a special z_T.

After sampling latent variable D, we sample latent variable z_T and z_t. According to our model, the probability will be decomposed into several conditions. For each word v of short text m which belongs to latent document d. The probability of specified z_T and z_t is as follows:

$P(z_{Ti}=k_T, z_{ti}=k_t|w, D, z_{T,-i}, z_{t,-i}, \gamma, \beta_T, \beta_t)=$

$$\begin{cases} (n_d^{z_T}+\beta_T^{z_T}-1)\dfrac{n_v^{z_T}+\gamma_v-1}{n^{z_T}+V\gamma-1} \\ (n_d^{z_T}+\beta_T^{z_T}-1)\dfrac{n_d^{z_T,z_t}+\beta_t^{z_t}-1}{n_d^{z_T}+K_{z_t}\beta_t^{z_t}-1}\dfrac{n_v^{z_T}+\gamma_v-1}{n^{z_T}+V\gamma-1} \\ (n_d^{z_T}+\beta_T^{z_T}-1)\dfrac{n_d^{z_T,z_t}+\beta_t^{z_t}-1}{n_d^{z_T}+K_{z_t}\beta_t^{z_t}-1}\dfrac{n_v^{z_T,z_t}+\gamma_v-1}{n^{z_T,z_t}+V\gamma-1} \end{cases}$$

Parameters η_z, θ_T, and θ_t are also be integrated out. We calculate $(n_d^{z_T}+\beta_T^{z_T}-1)\frac{n_v^{z_T}+\gamma_v-1}{n^{z_T}+V\gamma-1}$ if word v is assigned to root node. $n_d^{z_T}$ is the number of tokens in latent document d assigned to z_T. $n_v^{z_T}$ is the number of tokens equal to word v and assigned to topic z_T. n^{z_T} is the number of tokens assign to z_T of the document set. We calculate $(n_d^{z_T}+\beta_T^{z_T}-1)\frac{n_d^{z_T,z_t}+\beta_t^{z_t}-1}{n_d^{z_T}+K_{z_t}\beta_t^{z_t}-1}\frac{n_v^{z_T}+\gamma_v-1}{n^{z_T}+V\gamma-1}$ if word w is assigned to a super-topic. $n_d^{z_T}$ is the number of tokens assign to z_T and z_t in document d. We calculate $(n_d^{z_T}+\beta_T^{z_T}-1)\frac{n_d^{z_T,z_t}+\beta_t^{z_t}-1}{n_d^{z_T}+K_{z_t}\beta_t^{z_t}-1}\frac{n_v^{z_T,z_t}+\gamma_v-1}{n^{z_T,z_t}+V\gamma-1}$ if word v is assigned to a sub-topic. $n_v^{z_T,z_t}$ is the number of tokens equal to word v and assigned to topic z_T and z_t. n^{z_T,z_t} is the number of tokens assigned to topic z_T and z_t.

According to the inference equations above, we use Gibbs sampling method to sample the topic tree. Firstly, for each short text m, we sample long document D according to the probability $P(D_i=d|D_{-i},w, z_T, z_t, \alpha, \beta_T, \beta_t)$. If we sample a long document d, then we will calculate the topic tree next step. For each word in m, we firstly sample this word to the root topic or not. If root topic is sampled, we will turn to the next word. If the root topic is sampled, that means this word belongs to child topics. So we sample a topic from the child topics of the root topic. But if no child topic is sampled, that means this word should sample topics from leaf topics. So we sample a topic from leaf topics. All these sampling probabilities are following $P(z_{Ti}=k_T, z_{ti}=k_t|w, D, z_{T,-i}, z_{t,-i}, \gamma, \beta_T, \beta_t)$. The probability of not sampling root topic is 1- $P(z_{Ti}=k_T, z_{ti}=$

$k_t|w, D, z_{T,-i}, z_{t,-i}, \gamma, \beta_T, \beta_t$). And the probability of not sampling child topics is 1-$\sum P(z_{Ti} = k_T, z_{ti} = k_t|w, D, z_{T,-i}, z_{t,-i}, \gamma, \beta_T, \beta_t)$. $\sum P(z_{Ti} = k_T, z_{ti} = k_t|w, D, z_{T,-i}, z_{t,-i}, \gamma, \beta_T, \beta_t)$ is the summary of probabilities of child topics.

4. EXPERIMENTAL RESULTS

In this section, we first introduce the datasets in our experiment and the methods for comparison. Then we describe our evaluation with two metrics.

4.1. Datasets

We adopt four real world datasets to analyze performance. The statistics of these datasets are listed in Table 1. #Documents means the number of short texts. Dictionary size means the number of word. Avg. text size means the average length of short texts.

Table 1. Statistics of datasets.

DataSet	#Documents	Dictionary size	Avg. text size
News Titles	32,603	25,973	18
DBLP	628,363	51,799	7
Tweets	483,552	46,732	6
Reddit	494,704	50,276	7

We minimally pre-processed these datasets by removing stopwords and words that occur only once. We briefly describe them as follows:

News Titles:

This dataset of news titles is collected from RSS feeds of three popular newspaper websites (nyt.com, usatoday.com, reuters.com). There are 32k news titles across 7 categories (Sport, Business, U.S., Health, Sci\&Tech, World, and Entertainment). The description and the titles of news are combined as one short text.

DBLP:

This dataset consists of academic paper titles of computer science. These data are from DBLP, a bibliography website for computer science publications. We obtained 600k paper titles from DBLP database.

Tweets:

This dataset is collected from Twitter website. The content of tweets in Twitter is very informal. So we collected tweets of a specific area. This dataset contains 400k tweets, all about the 2016 United States elections.

Reddit News:

This dataset consists of news titles collected from Reddit. There are 500k news collected from 2008 to 2016.

4.2. Methods & Parameter Settings

In this section, we introduce the methods we implemented for comparison.

hPAM:

We implement this method as the base method. The model of this method suggests that each document has a distribution over the whole tree. The height of the tree is designed by the model. In this experiment, we apply the original model. So the height of the tree is constrained to be three.

nCRP:

This method is the typical model of a tree structure. Its model suggests that each document has a distribution over a path of the tree. Although the height of the tree is decided automatically, we only adopt top three levels of the tree for comparison.

hvHDP:

This method is the state of the art method of DAG structrue model. This model samples every level as an HDP. So this method is proposed under the framework of hHDP. This model not only samples topics at leaf nodes but also sample at non-leaf nodes.

For every method in this experiment, we generate a three-evel tree. For every topic, according to the probability distribution between words and topics, we select top 10 most probable words to represent topic content. Then, we search the PMI score of word co-occurrence between five words. The median PMI score is selected as the representation PMI score of the topic. Last, we calculate the avarage PMI score of the whole tree as the final coherent score.

The parameter setting of our method is as follows. Our method shPAM generates a three level tree with 1 root topic, 5 super-topics, and 10 sub-topics. The hyper-parameter β_T and β_t are all set to be 0.1. Hyper-parameter α is set to be 0.1 and γ is set to be 0.01. The number of latent document is set to be 2000. The parameters of other methods are set according to their papers. Method nCRP and hHDP do not need artificially setting the number of topics of each level. For nCRP, set γ =1.0, m= 100, η= 0.1 and π= 0.5. For hvHDP, set H=0.5, α=10, and λ=1.0. Method hPAM is also set to generate a tree with 5 super-topics and 10 sub-topics. Then set α=0.1, β= 0.01 and γ=10.

4.3. Evaluation Measures

Topic coherence is a common metric to evaluate topics. Here we use PMI score to get topic coherence. This metric needs auxiliary data to calculate PMI score. We chose the latest dump of Wikipedia articles as auxiliary data which contains 5 million documents and 14 million words of vocabulary. Firstly, we build a sliding window of 10 topics. Then we use this sliding window to get word co-occurrence information. Then we calculate PMI score according to the equation: PMI (w1,w2)=$\frac{P(w1,w2)}{P(w1)P(w2)}$. Here, P(w1,w2) means word w1 and word w2 appear in the same sliding window. P(w1) and P(w2) are marginal probabilities. So, according to PMI scores calculated for Wikipedia, we calculate the PMI score of each topics for hierarchical topic models. In our experiment, we choose top 10 words of each topic, and calculate the average PMI score and then calculate the average PMI score of all topics.

4.4. Topic Coherent Evaluation

In this section we show the result of the PMI score and analyse this result.

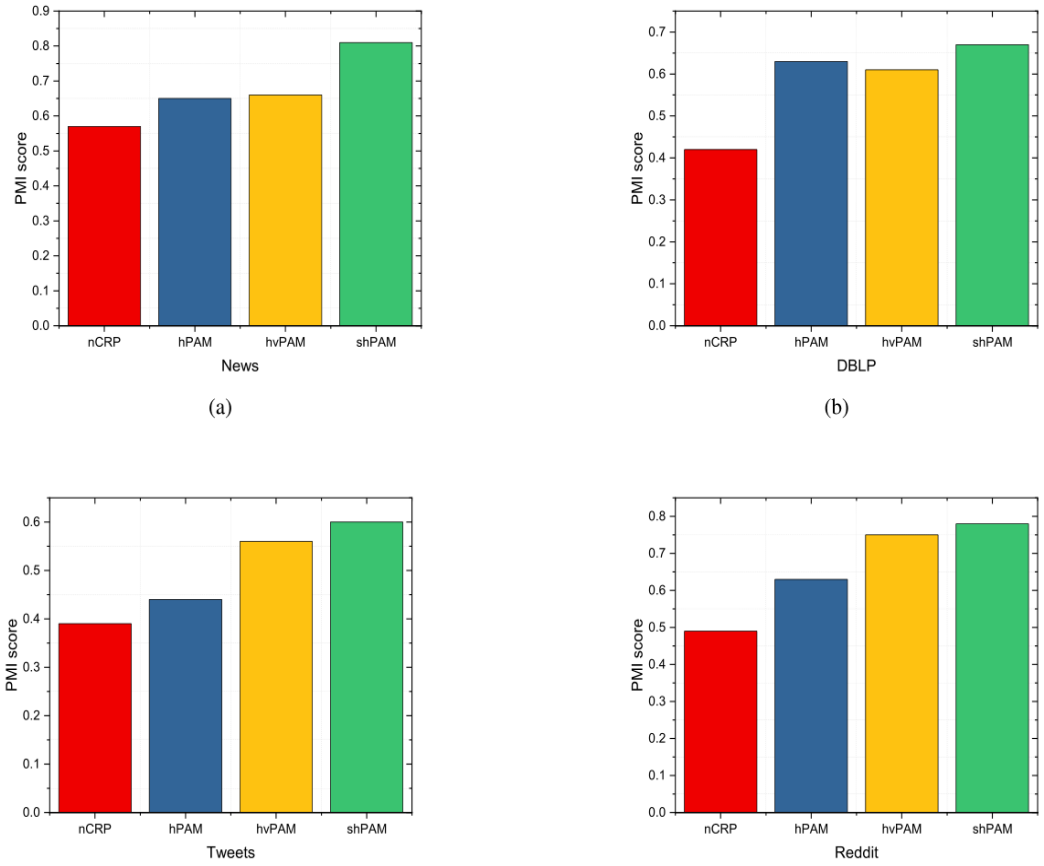


Figure 2. PMI score of 4 datasets

The results of PMI score can be seen in Figure 2. If the topic is more coherent, the PMI score will be higher. Method nCRP has the most un-coherent topics. This model supposes that one document corresponds to a path in the tree. But for short texts, co-occurrence will be more sparsity in subtrees. So the performance of this model is the poorest. Method shPAM, hvHDP, and hPAM all generate a DAG model, which can be seen outperforms tree structured model. Method hPAM is the baseline of DAG model. We can find our method shPAM outperforms it. The state-of-the-art method hvHDP is better than hPAM on dataset News, Tweets, and Reddit. Method hvHDP can provide a more coherent result by automatically changing the number of topics. But these methods cannot provide sufficient word co-occurrence. So the result shows that they all suffer from the sparsity of word co-occurrence. The result of our method shows that by aggregating short texts, we successfully incorporate additional word co-occurrence information into topic model. With more word co-occurrence, our model outperforms the other methods on all datasets.

5. CONCLUSIONS

In this paper, we propose a self-aggregated hierarchical topic model, especially for short texts. Hierarchical models for short texts will suffer from lacking word co-occurrence and general-special word relations. By incorporating long documents as latent variables, our model

aggregates short texts into long documents. Then long documents bring additional word co-occurrence and additional general-special word relations. The experiment on several real-world short texts corpus shows that our model can construct a hierarchy with more coherent topics than the state-of-the-art models. In future works, we will incorporate additional semantic information into hierarchy models such as short texts embeddings. Embeddings information will be helpful to overcome lacking word co-occurrence and general-special word relations.

REFERENCES

[1] Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018.

[2] Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM (JACM), 57(2), 1-30.

[3] Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2014). Nested hierarchical Dirichlet processes. IEEE transactions on pattern analysis and machine intelligence, 37(2), 256-270.

[4] Mimno, D., Li, W., & McCallum, A. (2007, June). Mixtures of hierarchical topics with pachinko allocation. In Proceedings of the 24th international conference on Machine learning (pp. 633-640).

[5] Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013, July). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 889-892).

[6] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

[7] Yu, D., Xu, D., Wang, D., & Ni, Z. (2019). Hierarchical topic modeling of Twitter data for online analytical processing. IEEE Access, 7, 12373-12385.

[8] Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003, December). Hierarchical topic models and the nested Chinese restaurant process. In NIPS (Vol. 16).

[9] Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).

[10] Kim, J. H., Kim, D., Kim, S., & Oh, A. (2012, October). Modeling topic hierarchies with the recursive chinese restaurant process. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 783-792).

[11] Ahmed, A., Hong, L., & Smola, A. (2013, May). Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In International Conference on Machine Learning (pp. 1426-1434). PMLR.

[12] Blundell, C., Teh, Y. W., & Heller, K. A. (2012). Bayesian rose trees. arXiv preprint arXiv:1203.3468.

[13] Song, Y., Liu, S., Liu, X., & Wang, H. (2015). Automatic taxonomy construction from keywords via scalable bayesian rose trees. IEEE Transactions on Knowledge and Data Engineering, 27(7), 1861-1874.

[14] Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., & Han, J. (2013, August). A phrase mining framework for recursive construction of a topical hierarchy. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 437-445).

[15] Zhang, Y., Mao, W., & Zeng, D. (2015, November). Constructing Topic Hierarchies from Social Media Data. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1015-1018). IEEE.

[16] Zuo, Y., Li, C., Lin, H., & Wu, J. (2021). Topic Modeling of Short Texts: A Pseudo-Document View with Word Embedding Enhancement. IEEE Transactions on Knowledge and Data Engineering.

AUTHORS

Yue Niu received the B.E. degree in software engineering from Central South University of China. He is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include text mining and machine learning.

HONGJIE ZHANG received the B.E. degree in software engineering from the Chongqing University. He is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include deep reinforcement learning, distributed system and cloud computing.

LEVERAGING OF WEIGHTED ENSEMBLE TECHNIQUE FOR IDENTIFYING MEDICAL CONCEPTS FROM CLINICAL TEXTS AT WORD AND PHRASE LEVEL

Dipankar Das and Krishna Sharma

Department of Computer Science & Engineering, Jadavpur University, India

ABSTRACT

Concept identification from medical texts becomes important due to digitization. However, it is not always feasible to identify all such medical concepts manually. Thus, in the present attempt, we have applied five machine learning classifiers (Support Vector Machine, K-Nearest Neighbours, Logistic Regression, Random Forest and Naïve Bayes) and one deep learning classifier (Long Short Term Memory) to identify medical concepts by training a total of 27.383K sentences. In addition, we have also developed a rule based phrase identification module to help the existing classifiers for identifying multi- word medical concepts. We have employed word2vec technique for feature extraction and PCA and T- SNE for conducting ablation study over various features to select important ones. Finally, we have adopted two different ensemble approaches, stacking and weighted sum to improve the performance of the individual classifier and significant improvements were observed with respect to each of the classifiers. It has been observed that phrase identification module plays an important role when dealing with individual classifier in identifying higher order n-gram medical concepts. Finally, the ensemble approach enhances the results over SVM that was showing initial improvement even after the application of phrase based module.

KEYWORDS

Medical Concepts, Phrase Identification, Ensemble, Machine Learning.

1. INTRODUCTION

In recent trends of digital platforms, people in general are relying on electronic data because the number of active internet users is increasing in medical domain[11]. It is very necessary to develop a state of the art tool to extract medical phrases from raw unstructured text.

We have developed a structured dataset of bio-medical concepts by manually annotating each and every term as either medical or not by collecting huge amount of raw data from the web archives. The training data consists of 27383 sentences while 7283 sentences are available as test data.

In the present work, we have developed a model that identifies medical concepts from texts as well as helps medical practitioners as well as novice users to deal with unstructured data. One instance of input and output of our model is shown as follows.

[1] http://www.nbcnews.com/id/3077086/t/more-people-search-health-online/

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.161-173, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111213

Input: *Amlodipine is used with or without other medications to treat high blood pressure*.

Output: ***Amlodipine_MC*** *is_O used__O with__O or_O without_O* ***[other medications] MC*** *to__O* ***treat__MC*** ***[high blood pressure]__MC .*** *_O .*

In the above example, the words (phrases) tagged with "__MC" are medical terms and the words that are tagged with "__O" are non-medical terms. The phrases are separated with brackets "[ ]". It has been observed that the presence of non-medical words also invokes the sense of a medical concept. For example, the words in italic are non-medical words whereas their appearance along with medical words forms a phrase level medical concept (e.g., "*Rat* Fever", "*Indian* Medical *Association*" etc.). For this reason, phrase identification module plays an important role and some set of rules are defined by considering medical as well as linguistic features. Moreover, support and confidence are also measured in order to identify the best possible phrase identification rules to tag multi-word medical concepts. Performances of the individual classifier before and after applying phrase identification are less while comparing the performance of the ensemble approach.

Finally, we have applied an ensemble approach to combine multiple classifiers to predict better than that of the individual classifier. The evaluation result shows that the ensemble approaches outperform other classifiers. We have applied two ensemble approaches i.e. stacking and weighted sum. Stacking helps to identify unigram medical concepts whereas weighted sum out performs multiword n-grams where n lies between 2 to 5**.**

The rest of the paper is organized as follows. The literature survey on extracting medical entities by machine learning classifiers is discussed in Section 2. The dataset preparation is discussed in Section 3 whereas machine learning and deep learning frameworks are described in Section 4. The phrase identification module is described in Section 5 followed by its evaluation results over ML approaches as discussed in Section 6. In contrast to ML and DL, Section 7 illustrates the implications of ensemble approaches and Section 8 highlights the feature selection strategies for improving results along with critical observations. Finally, Section 9 concludes the paper by mentioning future tasks.

2. RELATED WORK

Biomedical information extraction from the unstructured data is considered as one of the emerging challenges in the research field of NLP. Hence, a domain specific lexicon has become an essential component for converting a structured corpus from the unstructured corpus. Also, it helps in extracting the subjective and conceptual information related to medical concepts from the corpus.

Various researchers have tried to build various ontologies and lexicons such as UMLS, SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), MWN (Medical WordNet), SentiHealth, and WordNet of Medical Events (WME 1.0 and WME 2.0) etc. in the domain of healthcare.

UMLS helps to enhance the access to biomedical literature by facilitating the development of computer systems that understand biomedical language (Bodenreider, 2004). SNOMED-CT is a standardized, multilingual vocabulary that contains clinical terminologies and assists in exchanging the electronic healthcare information among physicians (Donnelly, 2006).

Furthermore, Fellbaum and Smith (2004) proposed Medical WordNet (MWN) with two sub-

networks e.g., Medical FactNet (MFN) and Medical BeliefNet (MBN) for justifying the consumer health. The MWN follows the formal architecture of the Princeton WordNet (Fellbaum, 1998). On the other hand, MFN aids in extracting and understanding the generic medical information for non-expert groups whereas MBN identifies the fraction of the beliefs about the medical phenomena (Smith and Fellbaum, 2004). Their primary motivation was to develop a network for medical information retrieval system with visualization effect.

Being in the similar trends, SentiHealth lexicon was developed to identify the sentiment of the medical concepts (Asghar et al., 2016; Asghar et al., 2014). In recent times, WME 1.0 and WME 2.0 lexicons were designed to extract the medical concepts and their related linguistic and sentiment features from the corpus (Mondal et al., 2016; Mondal et al., 2018).

These mentioned ontologies and lexicons assist in identifying the medical concepts and their sentiments from the corpus but unable to provide the complete knowledge of such concepts. Hence, in the current work, we are motivated to design a full-fledged lexicon in healthcare which provides the linguistic and knowledge-based features together for the medical concepts.

3. DATA PREPARATION

A total of 170 medical e-books of various sub-domains such as anatomy, internal, medicine, physiology, biochemistry etc. were collected from various web archives. Such books are mainly recommended for medical degree courses. Some of the books are text books[2], some books are medical encyclopedia[3], and a few are medical dictionaries[4]. We have extracted texts from the pdf files of all such books using open source tika[5] python library.

Finally, we have trained a word2vec word embedding model (Embedding size ~ 100) using these texts. We have used gensim[6] python library for training purpose. This large collection of text is used only for training our own word embedding whereas we have selected only a part of these texts for training and test purposes of the machine learning and deep learning classifiers, separately.

On the other hand, we collected a total of 34666 sentences from a medical dictionary[7]. These sentences are split into 27383 sentences for training and 7283 sentences for test purposes. The training set contains 498734 words whereas test set contains 130662 words, respectively. We have mentioned the brief details of our training and test data. In this Table 1, the statistics denote for medical words / phrases only.

[2] https://medicostimes.com/all-mbbs-books-pdf/
[3] Gale encyclopedia vol 1 to 5
[4] Dictionary of Medical Terms 4th Ed.- (Malestrom) and Black's medical dictionary etc.
[5] https://pypi.org/project/tika
[6] https://radimrehurek.com/gensim/models/word2vec.html
[7] BLACK'S MEDICAL DICTIONARY 41ST EDITION

Table 1. Statistics of the dataset

Dataset	# of N in N-gram					# phrases
	N=1	*N=2*	*N=3*	*N=4*	*N=5*	
Training	43734	41273	14904	4134	1247	134826
Test	11242	12245	3889	1048	334	29534
Total	54976	51518	18793	5182	1581	164360

4. SYSTEM FRAMEWORK

We have used five machine learning models followed by one deep learning model. We have used SVM (degree of SVM polynomial kernel is 3, and C=1.0.), K-NN (K=4), Logistic Regression, Gaussian Naïve Bayes and Random Forest algorithms for developing our machine learning framework.

We have applied these 5 machine learning classifiers to explore a comparative study among their performances with respect to the classification of medical concepts. Apart from that, we have selected multiple classifiers because we wanted to enhance the performance of classification framework by applying ensemble technique. As the classifiers require features for learning, we have used word embedding to convert word to feature vector and employed as features.

For machine learning classifiers, we have used scikit-learn[8] python library and for our deep learning framework using LSTM (Long and Short Term Memory) model, we have used keras[9] python library. In the LSTM, we have used time distributed character embedding with output dimension 20. In this layer, we have used LSTM unitof 64 with recurrent dropout=0.1. In the next layer, we have used LSTM unit of 256 with recurrent dropout=0.1 and in the last layer, we have employed a dense layer with *softmax* activation function. In this model, we have used *adam* optimizer with *binary_crossentropy* loss function.

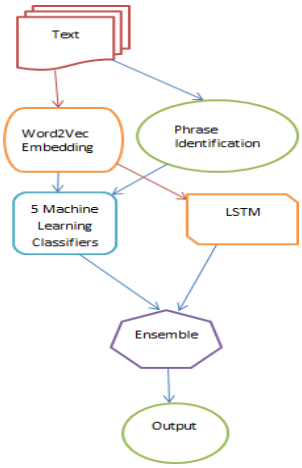


Figure 1: System Framework

[8] https://www.scikit-learn.org
[9] https://www.keras.io

We have discussed various steps for concept identification from medical texts. If we summarize all the process and understand the steps one by one, we have to look into the following diagram. In the above diagram, texts are sent to word2vec model for feature extraction that extracts the words and their features. The same text is sent to phrase identification module and it returns a list of words and phrases according to the sequence of the input text. Each word and its corresponding vectors were extracted from the text are sent to machine learning classifiers. If our classifiers observe a single word, it predicts whether it is medical term or not and if it does not see a multi word, it predicts such that if a multiword expression contains at least one medical term or notand declares this as a medical phrase. This same text with word2vec word embedding is sent to LSTM module. We send the outputs of the classifiers to ensemble module to increase the performance. After ensemble, we receive the final output.

However, these classifiers are not good enough with sequential data and thus unable to classify multi-word or phrase level medical concepts. In order to identify such phrase level medical concepts, we have developed a rule based phrase identification module for our task. The phrase level module helps in machine learning classifiers whereas in case of developing our deep learning framework using LSTM model with character embedding, we did not employ phrase identification module. We have also compared the performance between LSTM and the 5 machine learning classifiers with phrase identification.

5. PHRASE IDENTIFICATION

As the presence of non-medical words also invokes the sense of a medical concept, we have developed the rule based phrase identification module. Further, in order to handle sequential data using 5 machine learning classifiers, we have built this.

If we want to predict a medical phrase having a non-medical term as a whole, our classifier will predict that non-medical term as a medical whenever it occurs. E.g., **"***Indian Medical Association***"** is a medical phrase where the words, "*Indian*" and "*Association*" are not medical terms. However, if we want to predict this phrase as medical, our phrase identification module plays a vital role. We have used nltk[10] python library for phrase identification. The algorithm is as follows:

Step 1: *In the first pass, our algorithm will extract single and multiword medical concepts using phrase identification module.*

Step 2: *In the second pass, our classifier will predict all the single word expression and multiword such that if a multiword expression contains at least one medical term it predicts this as medical phrase*.

We have defined some phrase identification rules. Support and Confidence are measured in order to identify the best possible phrase identification rules to identify multi word medical concepts. The descriptions are given below.
In the above Table 2, we can observe that rule 1, 2, 3, 4, 10, 12, 14 are crucial to find multi word medical concepts and finally, we selected these rules only while

[10] https://www.nltk.org

applying into the framework of machine learning. In the next section, we have compared the performance of LSTM model with respect to each of the five machine learning classifiers and the importance of this phrase identification module is visible. Some example of the medical phrases with respect to each of the rule is given below.

Table 2: List of all phrase identification rules

RULES	**Support**	**Confidence**
RULE1 : {<JJ|NN|NNP><NNS>}	4523	0.258
RULE2{<VBP|VBN|NN|DT><NN>+}	7237	0.413
RULE3: {<RB><VBD|VB><NN>}	485	0.027
RULE4: {<DT><JJ>+<NN>}	3753	0.214
RULE5: {<DT><NN><JJ>}	26	0.0014
RULE6: {<VB><IN><NN>}	38	0.0021
RULE7: {<VB><TO><NN>}	8	0.0004
RULE8: {<NN><IN><VBG>}	84	0.004
RULE9: {<NN><CC><NN><VBZ>}	62	0.003
RULE10:{<NN.><NN.><NN>*}	430	0.024
RULE11: {<DT><RB><JJ><NN>*}	87	0.004
RULE12: {<CD>*<NN><IN><NN>}	384	0.021
RULE13: {<NNP><NN>+}	88	0.005
RULE14: {<RB>*<CD><NNS>}	311	0.017

Table 3: Phrase identification rules with examples

RULES	**Examples**
RULE1 : {<JJ|NN|NNP><NNS>}	Severe symptoms
RULE2{<VBP|VBN|NN|DT><NN>+}	The liver
RULE3: {<RB><VBD|VB><NN>}	Significantly lower cholesterol
RULE4: {<DT><JJ>+<NN>}	The tympanic membrane
RULE5: {<DT><NN><JJ>}	The autonomic nervous System
RULE6: {<VB><IN><NN>}	Lack of oxygen
RULE7: {<VB><TO><NN>}	Leads to death
RULE8: {<NN><IN><VBG>}	difficulty in breathing
RULE9: {<NN><CC><NN>}	Anoxia and hypoxia
RULE10:{<NN.><NN.><NN>*}	Catecholamine substances
RULE11: {<DT><RB><JJ><NN>*}	A potentially life-threatening condition
RULE12: {<CD>*<NN><IN><NN>}	Encyclopedia of medicine
RULE13: {<NNP><NN>+}	X Chromosome
RULE14: {<RB>*<CD><NNS>}	22 autosomes

6. EVALUATION

The test dataset consists of 7283 sentences (130662 words) manually annotated. We have analyzed 5 traditional ML algorithms (SVM, K-NN, LR, Naïve Bayes, Random Forest), and we have shown that these classifiers can also perform well in sequential data while using phrase identification module. We have used one deep learning (LSTM with character embedding) for classification to avoid phrase identification. As LSTM performs better in case of the sequential data by default, therefore, we did not apply phrase identification module into it. We have evaluated the performance of every classifier in phrase level. After evaluation we have increased our model's performance using ensemble method.

Table 4: Performance metrics of the classifiers

Classifiers	N in N-gram	**Precision**	**Recall**	**F1-Score After** phrase identification	**F1-Score Before** phrase identification
SVM	1	0.88	**0.95**	**0.92**	**0.92**
	2	**0.91**	0.93	**0.92**	**0.81**
	3	0.87	0.92	0.90	0.78
	4	0.72	0.94	0.81	0.60
	5	0.55	0.80	0.65	0.00
RandomForest	1	**0.77**	0.73	0.75	0.75
	2	**0.77**	0.79	**0.78**	0.72
	3	0.65	0.80	0.72	0.74
	4	0.39	**0.84**	0.53	0.41
	5	0.22	0.60	0.32	0.00
Naïve Bayes	1	0.63	**0.87**	0.73	0.73
	2	**0.78**	0.85	**0.81**	0.74
	3	0.67	0.85	0.75	0.60
	4	0.51	0.85	0.64	0.55
	5	0.35	0.73	0.75	0.00
Logistic Regression	1	0.71	0.72	0.71	0.71
	2	**0.73**	0.77	**0.75**	**0.65**
	3	0.60	0.75	0.66	0.59
	4	0.35	**0.78**	0.49	0.38
	5	0.20	0.53	0.29	0.0
KNN	1	0.88	**0.95**	**0.92**	**0.92**
	2	**0.91**	0.93	**0.92**	**0.81**
	3	0.87	0.92	0.90	0.76
	4	0.72	0.94	0.81	0.70
	5	0.55	0.80	0.65	0.00
LSTM	1	0.87	0.91	0.89	NA
	2	0.90	0.92	0.91	NA
	3	0.83	0.78	0.80	NA
	4	0.69	0.65	0.67	NA
	5	0.58	0.60	0.59	NA

From the above Table 4, we can observe that SVM, KNN and LSTM have performed well among all the other classifiers. From the last two columns (F1-Score after and before phrase identification), we can conclude that phrase identification plays a vital role in multi-word medical concept. It is also observed that the performance is decreased while predicting higher order N-grams. As phrase identification is not used in LSTM, performance analysis of before and after phrase identification of LSTM is not applicable. It is also noticed that SVM and K-NN with phrase identification rules perform better than LSTM. It means we can conclude that phrase identification is a key task of medical concept identification and classification.

7. ENSEMBLE APPROACH

In conventional machine learning, ensemble is a technique that uses multiple learning algorithms to obtain better performance which could not be obtained from any of the single learning algorithm alone. In this paper, we have used one type of ensemble approach i.e. “Weighted_Sum”. We will discuss about the performance gain as follows.

We have used six different classifiers and observed that three classifiers (e.g., SVM, LSTM and KNN) performed well and rest of the three (Random Forest, Naïve Bayes and Logistic Regression) performed moderate. As we tried to increase the overall performance, we finally selected SVM, KNN, LSTM, RF as the top performers and therefore ensemble them to improve the performance of our system.

We used weighted sum approach for ensemble and have given higher weight to the classifiers that obtain higher accuracy. Similarity in results between a pair of classifiers with respect to specific n-grams is also observed. Our motivation is to combine such output and predict better. The weighted sum is calculated as follows. Suppose, we have *n* number of classifiers and their outputs are α_1, α_2……..α_{n-1}, α_n. If we have assigned certain weights for each of the classifiers such as ω_1, ω_2……ω_{n-1}, ω_n, then the weighted sum for the output of the classifiers is If $\sum (\alpha_i * \omega_i) > £$ (£ is some threshold value), we classify it as **"MC"** class and otherwise, we classify it as "**O**" class.

In our approach, we have started by employing all the six classifiers and let the output of the classifiers are: *svm, knn, lr, lstm, nb* and *rf,* respectively. We have compared multiple possible weightages of all the classifiers and compared the F1-Score of all the combinations. In the following Table, we have given some instances of the combination of weights with respective £.

Table 5.1: Instances of the combination of weights When £=0.4

When £=0.4	**SVM**	**KNN**	**LSTM**	**LR**	**F1**
	0.25	0.25	0.25	0.25	0.91
	0.25	0.30	0.30	0.15	0.92
	0.10	0.25	0.30	0.35	0.82
	0.30	0.30	0.30	0.10	0.93
	0.35	0.30	0.25	0.10	0.93

Table 5.2: instances of the combination of weights When £=0.5

SVM	**KNN**	**LSTM**	**LR**	**F1**
0.25	0.25	0.25	0.25	0.91
0.25	0.30	0.30	0.15	0.88
0.10	0.25	0.30	0.35	0.84
0.30	0.30	0.30	0.10	0.91
0.35	0.30	0.25	0.10	0.92

Table 5.3: instances of the combination of weights When £=0.6

SVM	**KNN**	**LSTM**	**LR**	**F1**
0.25	0.25	0.25	0.25	0.92
0.25	0.30	0.30	0.15	0.91
0.10	0.25	0.30	0.35	0.82
0.30	0.30	0.30	0.10	0.92
0.35	0.30	0.25	0.10	0.94

Table 5. 4: instances of the combination of weights When £=0.7

SVM	KNN	LSTM	LR	F1
0.25	0.25	0.25	0.25	0.93
0.25	0.30	0.30	0.15	0.93
0.10	0.25	0.30	0.35	0.84
0.30	0.30	0.30	0.10	0.94
0.35	0.30	0.25	0.10	0.96

\

Table 5.5: instances of the combination of weights When £=0.65

SVM	KNN	LSTM	LR	F1
0.25	0.25	0.25	0.25	0.88
0.25	0.30	0.30	0.15	0.83
0.10	0.25	0.30	0.35	0.80
0.30	0.30	0.30	0.10	0.89
0.35	0.30	0.25	0.10	0.89

From the above tables from 5.1 to 5.5, we can find that our optimal weights are $W = \{0.35, 0.3, 0.25, 0.1\}$ and optimal threshold value, $£ = 0.65$. From our observation, we have derived the following weighted sum ensemble equation.

$$\sum (\alpha_i * \omega_i) = .35*svm+0.3*knn+0.25*lstm+0.1*lr$$

If $\sum (\alpha_I * \omega_i) > 0.65$, we will classify it as "**MC**" class, otherwise, we classify it as "**O**" class. Using this approach, we have improved the performance of our classifiers. The performances of our classifiers after ensemble are shown in Table 6. We can observe that after ensemble, the performances of our classifiers have increased, especially for the multi-gram concept identification.

Table 6. Performance metrics after ensemble

Number of N in N-gram	Precision	Recall	F1-Score
0	0.98	0.99	0.99
1	0.90	0.96	0.93
2	0.92	0.96	0.94
3	0.91	0.94	0.93
4	0.89	0.94	0.91
5	0.86	0.86	0.86

8. FEATURE SELECTION

As mentioned earlier, we have used 100 length word2vec feature vector for learning. As the length is very large, we had to reduce the feature length. For this reason we conducted an ablation study. We have used PCA and *t-sne* for ablation study. We wanted to reduce the dimension from 100 to 20. The classification report is in the

following based on precision, recall and F1-Score. We have trained and tested on same data with new features. The performance matrices are as follows.

Table 7: Performance metrics after *t-sne*

Classifiers	Number of N in N-gram	Precision	Recall	F1-Score With phrase identification
SVM	1	0.83	**0.90**	**0.87**
	2	**0.82**	0.87	**0.85**
	3	0.80	0.88	0.84
	4	0.64	0.80	0.72
	5	0.50	0.72	0.59
Random Forest	1	**0.70**	0.68	0.67
	2	**0.69**	0.73	**0.72**
	3	0.59	0.76	0.67
	4	0.33	**0.81**	0.48
	5	0.22	0.58	0.31
Naïve Bayes	1	0.57	**0.81**	0.68
	2	**0.72**	0.80	**0.77**
	3	0.59	0.80	0.68
	4	0.51	0.81	0.63
	5	0.32	0.68	0.45
Logistic Regression	1	0.64	0.68	0.69
	2	**0.69**	0.72	**0.71**
	3	0.55	0.71	0.62
	4	0.30	**0.71**	0.42
	5	0.20	0.50	0.28
KNN	1	0.81	**0.90**	**0.85**
	2	**0.85**	0.89	**0.87**
	3	0.80	0.84	0.82
	4	0.67	0.77	0.78
	5	0.50	0.70	0.59
LSTM	1	0.81	0.87	0.84
	2	0.85	0.86	0.87
	3	0.79	0.74	0.77
	4	0.65	0.61	0.63
	5	0.50	0.51	0.51

Table 8: Performance metrics after PCA

Classifiers	Number of N in N-gram	Precision	Recall	F1-Score With phrase identification
SVM	1	0.85	**0.91**	**0.88**
	2	**0.84**	0.89	**0.86**
	3	0.81	0.88	0.84
	4	0.65	0.85	0.74
	5	0.50	0.76	0.61
Random Forest	1	**0.71**	0.69	0.68
	2	**0.71**	0.74	**0.73**
	3	0.61	0.78	0.69
	4	0.34	**0.80**	0.48
	5	0.22	0.58	0.31
Naïve Bayes	1	0.59	**0.82**	0.69
	2	**0.74**	0.81	**0.78**
	3	0.61	0.82	0.70
	4	0.51	0.81	0.63
	5	0.35	0.70	0.47
Logistic Regression	1	0.65	0.69	0.70
	2	**0.69**	0.72	**0.71**
	3	0.56	0.72	0.63
	4	0.31	**0.73**	0.44
	5	0.20	0.50	0.28
KNN	1	0.83	**0.91**	**0.86**
	2	**0.87**	0.90	**0.89**
	3	0.82	0.89	0.85
	4	0.70	0.90	0.79
	5	0.51	0.71	0.60
LSTM	1	0.82	0.88	0.85
	2	0.88	0.87	0.88
	3	0.80	0.77	0.78
	4	0.65	0.61	0.63
	5	0.51	0.55	0.53

In Table7, we have shown the performance of the classifiers after applying *t-sne* feature selection technique. After reducing the dimensions, we have seen that the performances were also reduced a bit. It means that we have lost some information after dimensionality reduction. Now, we have explored the performances of the individual classifiers after using PCA selection technique. We also completed a comparative study about PCA and *t-sne*. From Table 7 and Table 8, we can observe that performances have been reduced in both PCA *and t-sne*. However, *t-sne* performed better than PCA.

8.1. Observations

We have previously discussed that the presence of non-medical words also invokes the sense of a medical concept. We have used six machine learning classifiers. These classifiers are not good for phrase identification.

For example: **"World Health Organization"** is a medical phrase. In this phrase, "World" and "Organization" are not medical terms. If we want to label all the terms as medical, our classifiers predict "Indian" and "Association" as a medical term all the time where ever it will occur. For this reason, phrase identification plays a vital role. For understanding more, let's consider two sentences in the following

Sentence 1: *World Health Organization did not recommend Hydroxychloroquine as a medicine of Covid-19.*
Sentence 2: *World is in danger for a disease called Covid-19,*

In the two sentences, the word "world" has been used for two aspects. In first sentence "world" should be classified as medical and in the second sentence it should be classified as non-medical. But our ML classifiers will predict the word "world" as same (medical or non-medical) whenever it occurs. If we follow our method, it will correctly classify the two sentences. In the first sentence there is one phrase whereas in the second sentence, there is no phrase.

In the first pass:

Sentence 1: ***World Health Organization_PHRASE*** *did not recommend Hydroxychloroquine as a medicine of Covid-19.*

Sentence 2: *World is in danger for a disease called Covid-19.*

In the 2nd pass:

Sentence 1: *[World Health Organization]_MC did_O not_O recommend_O Hydroxychloroquine_MC as_O a_O medicine_MC of Covid-19_MC.*

Sentence 2: *World_O is_O in_O danger_O for_O a_O disease_MC called Covid-19_MC.*

In the first sentence, the phrase, "*World Health Organization*" contains a medical term "*Health*", for this reason "*World Health Organization*" becomes a medical term. In the second sentence there is no phrase. In this way we have dealt with two situations using traditional machine learning classifiers.

9. CONCLUSIONS

We have developed a module for concept identification in medical text. We have identified a phrase from a given text using some rules. We have created 6 types of binary classifiers to predict a word (phrase) is medical word (phrase) or not. We have analyzed the performances of these multiple classifiers. In our observation phrase identification module with SVM or K-NN performs better than LSTM. In this way we have shown the importance of phrase identification module. We have applied ensemble (Weighted sum) module for increasing accuracy. After all of these we have built a system which can identify medical concepts from unstructured medical plain text. In future, we are planning to integrate the model with chatbot for medical assistance.

ACKNOWLEDGEMENTS

The work is supported by the project "*Sevak- an Intelligent Indian Language Chatbot*" fundedby DST-SERB, MeiTY, Govt. of India.

REFERENCES

[1] Asghar Muhammad Z., S. Ahmad, M. Qasim, S. Rabail Zahra, and F. Masud Kundi. 2016. SentiHealth: creating health-related sentiment lexicon using hybrid approach. Springer-Plus, 5(1):1139.

[2] Asghar Muhammad Z ., A. Khan, F. M Kundi, M. Qasim, F. Khan, R. Ullah and I. U Nawaz. 2014. Medical opinion lexicon: an incremental model for mining health reviews. International Journal of Academic Research, 6(1):295–302.

[3] Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1):D267–D270.

[4] Donnelly, K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics, 121:279.

[5] Smith, B. and Fellbaum, C. 2004. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings of the 20th international conference on Computational Linguistics, page 371. Association for Computational Linguistics.

[6] Miller, G. and Fellbaum, C. 1998. Word-Net: An electronic lexical database.

[7] Mondal, A., D. Das, E. Cambria and S. Bandyopadhyay. 2016. WME: Sense, Polarity and Affinity based Concept Resource for Medical Events. Proceedings of the Eighth Global WordNet Conference, pages 242–246

[8] Mondal, A., D. Das, E. Cambria and S. Bandyopadhyay. 2018 WME 3.0: An Enhanced and Validated Lexicon of Medical Concepts Proceedings of the 9th Global WordNet Conference (GWC 2018), 10-16

MACHINE LEARNING AND DEEP LEARNINGTECHNOLOGIES

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 machine learning, which is the application of advanced deep learning techniques on big data. This paper aims to analyse some of the different machine learning and deep learning algorithms and methods, aswell as the opportunities provided by the AI applications in various decision making domains.

KEYWORDS

Artificial Intelligence, Machine Learning, Deep 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 avery 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 and deep 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 and deep learning algorithms and methods which can be applied to big data analysis, as well as the opportunities provided by the AI applications in various decision making domains.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.175-183, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111214

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:

- **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.
- **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.
- **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]:

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.

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, cross-selling strategies, customer segmentation, image and pattern recognition [6]. 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, k-means clustering, probabilistic clustering methods, and more [7].

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 [8]. 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]:

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 [9].

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 [10]. 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.

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 [11]. 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.

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 [12].

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 DEEP LEARNING

Deep learning is one of the foundations of artificial intelligence (AI), and the current interest in deep learning is due in part to the buzz surrounding AI. Deep learning techniques have improved the ability to classify, recognize, detect and describe – in one word, understand [13]. For example, deep learning is used to classify images, recognize speech, detect objects and describe content.

Several developments are now advancing deep learning:

- Algorithmic improvements have boosted the performance of deep learning methods.
- New machine learning approaches have improved accuracy of models.
- New classes of neural networks have been developed that fit well for applications like texttranslation and image classification.
- We have a lot more data available to build neural networks with many deep layers, including streaming data from the Internet of Things, textual data from social media, physicians notes and investigative transcripts [14].
- Computational advances of distributed cloud computing and graphics processing units have put incredible computing power at our disposal. This level of computing power is necessary to train deep algorithms.

At the same time, human-to-machine interfaces have evolved greatly as well. The mouse and the keyboard are being replaced with gesture, swipe, touch and natural language, ushering in a renewed interest in AI and deep learning [15].

3.1. How Deep Learning Works

Deep learning changes how you think about representing the problems that you're solving with analytics. It moves from telling the computer how to solve a problem to training the computer to solve the problem itself.

A traditional approach to analytics is to use the data at hand to engineer features to derive new variables, then select an analytic model and finally estimate the parameters (or the unknowns) of that model. These techniques can yield predictive systems that do not generalize well because completeness and correctness depend on the quality of the model and its features [16]. For example, if you develop a fraud model with feature engineering, you start with a set of variables, and you most likely derive a model from those variables using data transformations. You may end up with 30,000 variables that your model depends on, then you have to shape the model, figure out which variables are meaningful, which ones are not, and so on. Adding more data requires you to do it all over again.

The new approach with deep learning is to replace the formulation and specification of the model with hierarchical characterizations (or layers) that learn to recognize latent features of the data from the regularities in the layers [17]. The paradigm shift with deep learning is a move from feature engineering to feature representation. The promise of deep learning is that it can lead to predictive systems that generalize well, adapt well, continuously improve as new data arrives, and are more dynamic than predictive systems built on hard business rules. You no longer fit a model. Instead, you train the task.

Deep learning is making a big impact across industries. In life sciences, deep learning can be used for advanced image analysis, research, drug discovery, prediction of health problems and disease symptoms, and the acceleration of insights from genomic sequencing. In transportation, it can help autonomous vehicles adapt to changing conditions [18]. It is also used to protect critical infrastructure and speed response.

Most deep learning methods use neural networks architectures, which is why deep learning models are often referred to as deep neural networks. The term "deep" usually refers to the number of hidden layers in the neural network. Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150. Deep learning models are trained

by using large sets of labelled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

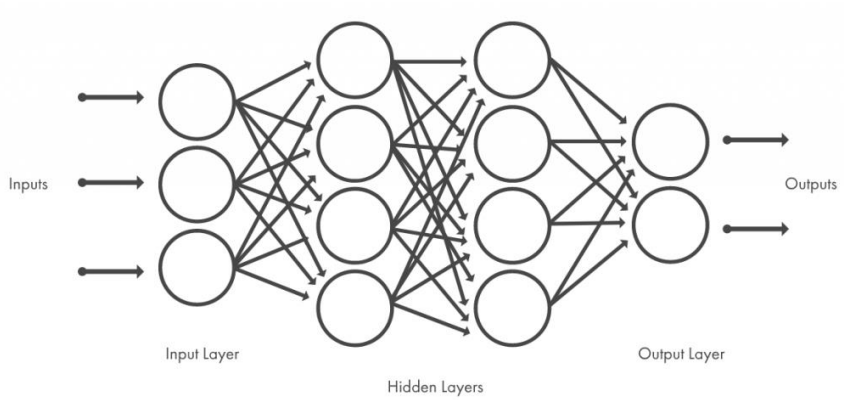


Figure 1. Neural networks, which are organized in layers consisting of a set of interconnectednodes. Networks can have tens or hundreds of hidden layers.

3.2. How Deep Learning Being Used

To the outside eye, deep learning may appear to be in a research phase as computer science researchers and data scientists continue to test its capabilities. However, deep learning has many practical applications that businesses are using today, and many more that will be used as research continues [19]. Popular uses today include:

Speech Recognition

Both the business and academic worlds have embraced deep learning for speech recognition. Xbox, Skype, Google Now and Apple's Siri, to name a few, are already employing deep learning technologies in their systems to recognize human speech and voice patterns.

Natural Language Processing

Neural networks, a central component of deep learning, have been used to process and analyse written text for many years. A specialization of text mining, this technique can be used to discover patterns in customer complaints, physician notes or news reports, to name a few.

Image Recognition

One practical application of image recognition is automatic image captioning and scene description. This could be crucial in law enforcement investigations for identifying criminal activity in thousands of photos submitted by bystanders in a crowded area where a crime has occurred. Self-driving cars will also benefit from image recognition through the use of 360-degree camera technology.

Recommendation Systems

Amazon and Netflix have popularized the notion of a recommendation system with a good chance of knowing what you might be interested in next, based on past behaviour. Deep learning can be used to enhance recommendations in complex environments such as music interests or clothing preferences across multiple platforms.

Recent advances in deep learning have improved to the point where deep learning outperforms humans in some tasks like classifying objects in images [20]. While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

1. Deep learning requires large amounts of labelled data. For example, driverless car development requires millions of images and thousands of hours of video.
2. Deep learning requires substantial computing power. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

When choosing between machine learning and deep learning, consider whether you have a high-performance GPU and lots of labelled data. If you don't have either of those things, it maymake more sense to use machine learning instead of deep learning. Deep learning is generally more complex, so you'll need at least a few thousand images to get reliable results. Having a high-performance GPU means the model will take less time to analyse all those images [21].

3.3. Deep Learning Opportunities and Applications

A lot of computational power is needed to solve deep learning problems because of the iterative nature of deep learning algorithms, their complexity as the number of layers increase, and the large volumes of data needed to train the networks.

The dynamic nature of deep learning methods – their ability to continuously improve and adapt to changes in the underlying information pattern – presents a great opportunity to introduce more dynamic behaviour into analytics [22]. Greater personalization of customer analytics is one possibility. Another great opportunity is to improve accuracy and performance in applications where neural networks have been used for a long time. Through better algorithms and more computing power, we can add greater depth.

While the current market focus of deep learning techniques is in applications of cognitive computing, there is also great potential in more traditional analytics applications, for example, time series analysis. Another opportunity is to simply be more efficient and streamlined in existing analytical operations. Recently, some study showed that with deep neural networks in speech-to-text transcription problems [23]. Compared to the standard techniques, the word- error-rate decreased by more than 10 percent when deep neural networks were applied. They also eliminated about 10 steps of data preprocessing, feature engineering and modelling. The impressive performance gains and the time savings when compared to feature engineering signify a paradigm shift.

Here are some examples of deep learning applications are used in different industries:

Automated Driving: Automotive researchers are using deep learning to automatically detect objects such as stop signs and traffic lights. In addition, deep learning is used to detect pedestrians, which helps decrease accidents.

Aerospace and Defence: Deep learning is used to identify objects from satellites that locate areas of interest, and identify safe or unsafe zones for troops.

Medical Research: Cancer researchers are using deep learning to automatically detect cancer cells. Teams at UCLA built an advanced microscope that yields a high-dimensional data set used to train a deep learning application to accurately identify cancer cells [24].

Industrial Automation: Deep learning is helping to improve worker safety around heavy machinery by automatically detecting when people or objects are within an unsafe distance of machines.

Electronics: Deep learning is being used in automated hearing and speech translation. For example, home assistance devices that respond to your voice and know your preferences are powered by deep learning applications.

3.4. How to Create and Train Deep Learning Models

The three most common ways people use deep learning to perform object classification are:

Training from Scratch

To train a deep network from scratch, you gather a very large labelled data set and design a network architecture that will learn the features and model. This is good for new applications, or applications that will have a large number of output categories. This is a less common approach because with the large amount of data and rate of learning, these networks typically take days or weeks to train [25].

Transfer Learning

Most deep learning applications use the transfer learning approach, a process that involves fine-tuning a pre-trained model. User can start with an existing network, such as AlexNet or GoogLeNet, and feed in new data containing previously unknown classes [26]. After making some tweaks to the network, user can now perform a new task, such as categorizing only dogs or cats instead of 10,000 different objects. This also has the advantage of needing much less data (processing thousands of images, rather than millions), so computation time drops to minutes or hours.

Feature Extraction

A slightly less common, more specialized approach to deep learning is to use the network as a feature extractor. Since all the layers are tasked with learning certain features from images, user can pull these features out of the network at any time during the training process [27]. These features can then be used as input to a machine learning model such as support vector machines (SVM).

4. CONCLUSIONS

So this study was concerned by understanding the inter-relationships between machine learning and deep learning, what frameworks and systems that worked, and how machine learning can impact the AI applications whether by introducing new innovations that foster advanced machine learning process and escalating power consumption, security issues and replacing human in workplaces. The advanced machine learning and deep 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] 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, Gurpreet Naul, 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, Vered Shenaar-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] Piero Mella, (2017). Intelligence and Stupidity – The Educational Power of Cipolla's Test and ofthe "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] Shi, Z., (2019). Cognitive Machine Learning. *International Journal of Intelligence Science*, 9, pp. 111-121.
[11] Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B., (2015). Human-Level Concept Learning through Probabilistic Program Induction. Science, 350, pp. 1332-1338.
[12] 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.
[13] Fukushima, K., Neocognitron: (1980). A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36,pp. 193-202.
[14] Lecun, Y., Bottou, L., Orr, G.B., et al., (1998). Efficient Backprop. Neural Networks Tricks ofthe Trade, 1524, 1998, pp. 9-50.
[15] Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning. The MIT Press,Cambridge.
[16] Fujii, K. (2018). Mathematical Reinforcement to the Minibatch of Deep Learning. Advances inPure Mathematics, 8, 307-320.
[17] Xuan, X., Peng, B., Wang, W. and Dong, J. (2019). On the Generalization of GAN Image Forensics. In: Chinese Conference on Biometric Recognition, Springer, Berlin, 134-141.
[18] Vaccari, C. and Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society,6, 1-13.
[19] Wang, F., Xing, L., Bagshaw, H., Buyyounouski, M. and Han, B. (2020). Deep Learning Applications in Automatic Needle Segmentation in Ultrasound-Guided Prostate Brachytherapy. Medical Physics.
[20] 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, 1995, pp. 419-457.
[21] 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.
[22] Wang, R., (2019). Research on Image Generation and Style Transfer Algorithm Based on Deep Learning. *Open Journal of Applied Sciences*, 9, pp. 661-672.
[23] Krizhevsky, A., Sutskever, I., Hinton, G.E., et al., (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems, 141, pp. 1097-1105.
[24] Long, J., Shelhamer, E., Darrell, T., et al., (2015). Fully Convolutional Networks for Semantic Segmentation. Computer Vision and Pattern Recognition, Boston, pp. 3431-3440.
[25] Noh, H., Hong, S., Han, B., et al., (2015). Learning Deconvolution Network for Semantic Segmentation. International Conference on Computer Vision, Santiago, pp. 1520-1528.
[26] Cheng, Z., Yang, Q., Sheng, B., et al., (2015). Deep Colorization. International Conference on Computer Vision, Santiago, 415-423.
[27] Mahendran, A. and Vedaldi, A., (2015). Understanding Deep Image Representations by Inverting Them. Computer Vision and Pattern Recognition, Boston, pp. 5188-5196.

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.

SKILLS MAPPING AND CAREER DEVELOPMENTANALYSIS USING ARTIFICIAL INTELLIGENCE

Yew Kee Wong

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

ABSTRACT

Artificial intelligence has been an eye-popping word that is impacting every industry in the world. With the rise of such advanced technology, there will be always a question regarding its impact on our social life, environment and economy thus impacting all efforts exerted towards continuous development. From the definition, the welfare of human beings is the core of continuous development. Continuous development is useful only when ordinary people's lives are improved whether in health, education, employment, environment, equality or justice. Securing decent jobs is a key enabler to promote the components of continuous development, economic growth, social welfare and environmental sustainability. The human resources are the precious resource for nations. The high unemployment and underemployment rates especially in youth is a great threat affecting the continuous economic development of many countries and is influenced by investment in education, and quality of living.

KEYWORDS

Artificial Intelligence, Conceptual Blueprint, Continuous Development, Human Resources, Learning and Employability Blueprint.

1. INTRODUCTION

Continuous development is defined as the development that meets the needs of the present without compromising the ability of future generations to meet their own needs [1]. The primary cause of the high unemployment rates is the inefficient education systems that fail to equip young people with the required skills for the labour market. In this research, we propose the use of artificial intelligence to enhance the relationship between education and employment.

Many studies were published on how to improve education curricula to enhance the employability of students; frameworks were designed to facilitate the work of teachers, mentors, career advisers and faculty to guide students through their career exploration and preparation. Numerous papers were published on the impact of artificial intelligence (AI) on education and its impact on employment. However it seems there is a gap in connecting the three important areas of research, 1: education for employment, 2: AI in education and, 3: AI in employment [2]. Further investigations are needed to evaluate and assess how AI can fit in the current learning and employability blueprint and to evaluate what can innovation and entrepreneurship bring to promote better education for employment systems.

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
Pp.185-192, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111215

2. USING AI TO BUILD A CONCEPTUAL BLUEPRINT

The study is assessing new blueprint for learning and employability and how AI can fit in and foster the process, so further experiments should be carried out to ensure the effectiveness of the blueprint and the accuracy of results of the AI application on the learning and employability process [3]. After reviewing literature regarding the impact of AI and its potential on both education and employment, as well as reviewing different education for employment blueprints, theories and case studies, this paper attempts to close the gap in the research related to specific scope which is the impact of AI on education for employment [4].

Young people can't find jobs. Yet employers can't find people with the required skill set. This mismatch between the supply and demand in the labour market might witness a bigger gap in the future with the growth of AI technologies. There are a few frameworks for education for employment or in other words "Learning and Employability" [5]. However the existing model didn't address the potential of AI whether in terms of deployment of such technology within the model or in terms of the implications of Al on the learning models or the employment models. So there is a need to find a practical frame for learning and employability that incorporate the advancements of AI to facilitate the university to work transition. This paper seeks to figure out the room for AI potentials through mapping innovative startups that embraced AI capabilities to play a role in the education for employment ecosystem.

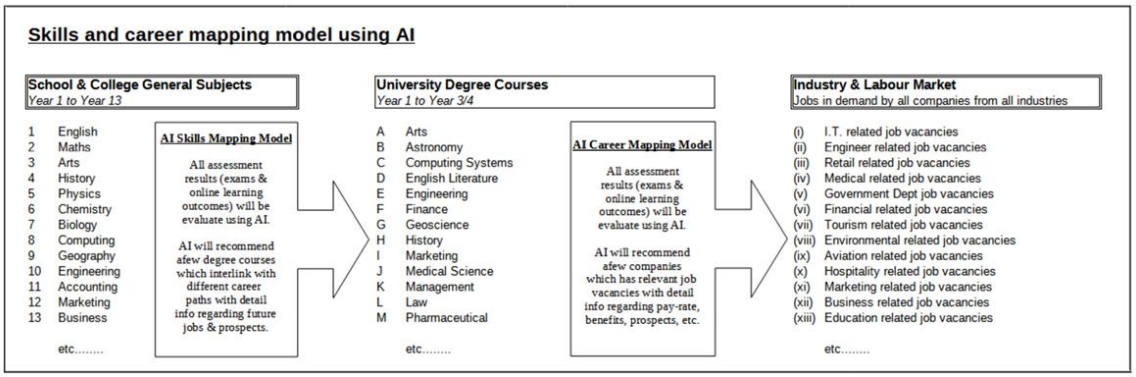


Figure 1. Skills and career mapping model using Al.

2.1. The Use of Al in Skills Mapping Model In this model

We proposed to use AI to streamline the skills required by various degree courses. This process significantly reduce the time for the student to decide on what degree subjects they can register for the university entrance. Furthermore, the model can also assist the student by presenting the detail information regarding the different career paths, current employers that are offering job related vacancies, pay-rate range, related benefits and other prospects [6].

Some students may not have a clear career path after they completed their college study and require further guidance and advice on choosing the appropriate degree course for their future career development [7]. In this evaluation process, the Al will use all the information (i.e. exam grades/marks, understanding level from online learning, etc..) provided by the student. Therefore, AI in this process can only provide advanced in-depth career roadmaps as recommendation for the student. The final decision making still rely on the student.

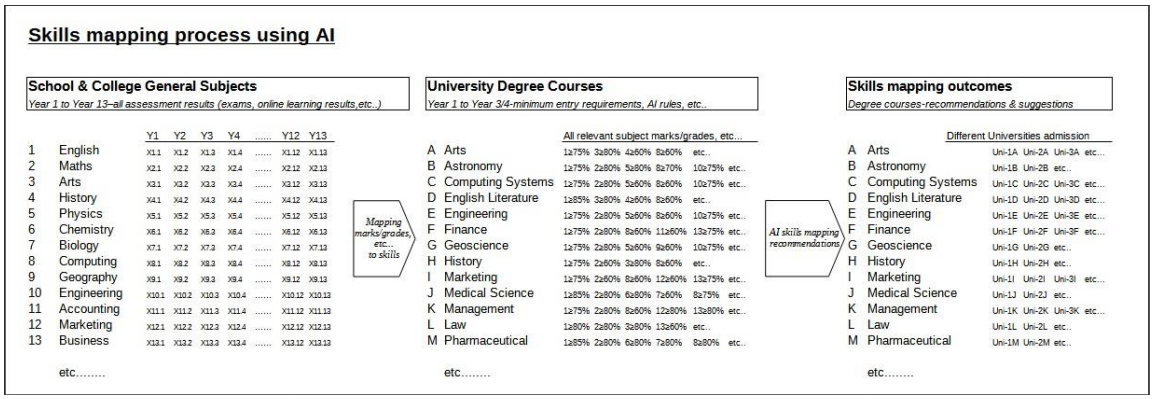


Figure 2. Skills mapping processing using AI.

2.2. The Use of AI in Career Mapping Model

With such employment concern, many studies refined such concern and the general consensus now is that AI will generate major transformations in the labour market [8]. According to many researchers, AI will create 2.3 million jobs in 2020, while eliminating 1.8 million and by 2025 AI related job creation will reach two million net-new jobs. Moreover, according to a new report from the World Economic Forum (WEF); 75 million jobs are estimated to be displaced, while 133 million new roles may emerge due to machines and algorithms [9]. The study has argued that this transition to technology should result in favourable unemployment that will allow human labour to better perform activities they were never able to do in their current heavyduty jobs. AI programs will probably be utilized for applications where hiring humans would be too expensive or really dangerous.

AI programs will take over computer tasks allowing humans to dedicate their time to other kinds of tasks including personal services. Service sector companies are optimistic about big data and enthusiastic about AI and robotics deployment as it will have direct impact on productivity improvement that eventually reflects on economic growth. On the other hand, it was realized that AI can positively impact employment if it is utilized properly within the business model [10]. AI uses in creating effective recruitment systems is seen as an inevitable opportunity to make best use of. Still this will stay challenging until firms management pay attention to the importance of allocating budgets to finance the required technology for hiring process.

Once the students graduated from the University, they can directly enter the labour market with the help from the AI career mapping model. The model will provide recommendations and information related to jobs, so that to the graduates can prepare for interviews and other job application processes, such as IQ & EQ tests, body check, etc.. This process can significantly reduce the amount of time the graduates need to search for jobs, interviews and other tedious job searching steps and at the same time also reduce the amount of time the employer canrecruit the appropriate personnel to fulfil the position and task required in the company, hence can indirectly improve the productivity rate of the company.

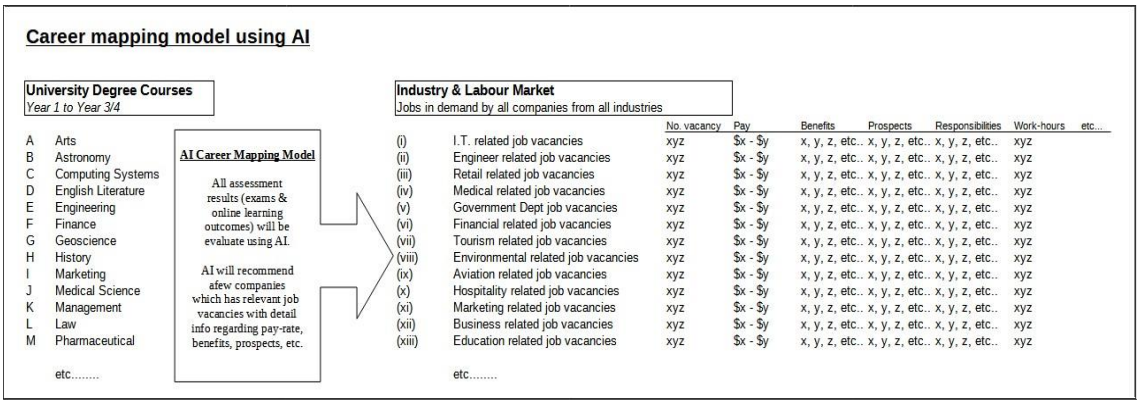


Figure 3. Career mapping processing using AI.

3. LEARNING AND EMPLOYABILITY BLUEPRINT

Aside from the impact of AI in creating new jobs, replacing jobs or even shifting the job and labour market, there are two global employment crises that already exist away from the implications of AI; high levels of youth unemployment and a shortage of talents who possess critical job skills. Mourshed, Farrell, & Barton [11] argued that if young people graduating from schools and universities, after exerting lots of efforts, cannot secure decent jobs and observe that sense of respect that comes with such degrees, society may witness outbreaks of anger or even violence. There is an information gap in what works and what does not in preparing young people during their school to employment transition. I summarized this information gap and it clearly shows there is a clear disconnect and misperception about youth job readiness from the point of view of employers vs youth vs educational institutions.

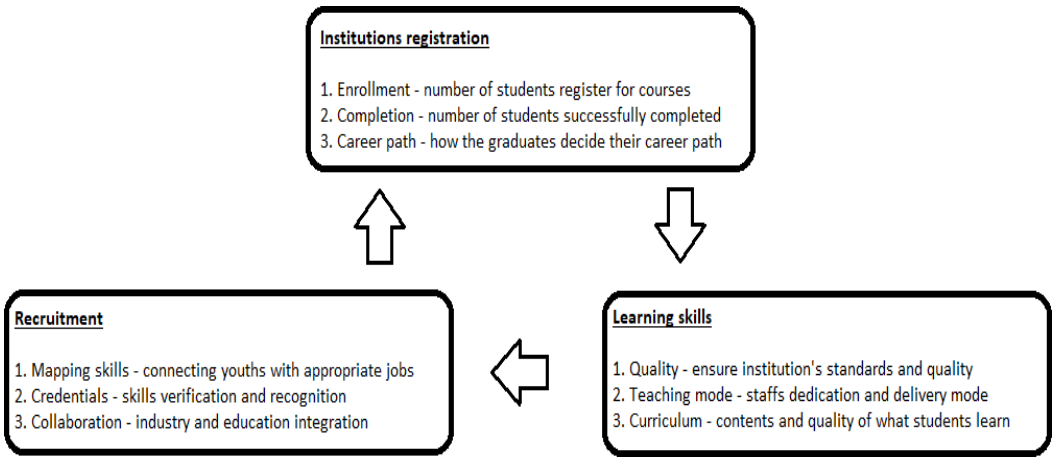


Figure 4. Blueprint for exploring the education to employment system.

3.1. Institutions Registration

A. Information Sharing

All institutions are recommended to develop a comprehensive occupations database and educational/training opportunities and provide information, advice and guidance to help job seekers to make decisions on learning, training and work. The comprehensive occupation database and website allows users to explore different career options including jobs profile, salaries, industry trends and offer webchats with career advisors beside their skills health check assessment that help users to find out what kind of jobs that best suits his/her skills. Users can also find training opportunities.

B. Dealing with Social Perception

It seems that a perception is widespread that getting a decent job with good salary requires being a college graduate. So this puts social pressure on youth to go to college and influence others' choice away from the vocational tracks [12]. Brunello and Rocco [13] argued that youth who graduated from vocational education have a higher likelihood of being not employed and with no education or training within the past 12 months. They also found that vocational education is associated with poorer labour market returns. This as a result impacted on the perception about vocational education.

C. Dealing with Education Affordability

Schultz [14] and Becker [15] introduced individual choice model of human capital investmentin which they presented individual's education choice as an investment decision. Individuals sacrifice economically in order to acquire knowledge, referred to as 'human capital', that will enable them to get better rewards in the future. If young people have no access to credit or savings, this may limit their choices and they will not be able to enrol in study.

3.2. Learning Skills

A. Effective Content and Curriculum Design

Mourshed et al. [11] proposed that in order to design relevant curriculum to the employers' requirements, close engagement between, industry leaders and educational providers is needed. Such engagement to succeed, intensive collaboration should exist while defining the core requirements on a very detailed level to ensure that the aspired learning outcomes will be achieved.

B. Effective Delivery Methods

Effective delivery requires still close engagement between employers and educational providers. Mourshed et al. [11] explored two main ways to do so - (1): Classrooms within workplaces. The common model to bring vocational and technical training within the workplaces is through internships or apprenticeships. (2): Workplaces within classrooms. Internships and apprenticeships are types of hands-on learning experiences that are most admired by students, however the number of opportunities are limited to accommodate certain capacities of students.

3.3. Recruitment

A. Assessment for Qualifications and Certifications

Finding a job is a painful process for job seekers. Job seekers strive to market their skills, but can't find a credible way to prove their talents, and Employers can't trust the educational degree as a main reference validating youth skills and knowledge. So both employers and candidates suffer in the hiring and talent acquisition process. One of the well known processes to show one's credentials and prove his skills and knowledge in a credible way is the international professional certifications such as PMP (Project Management Professional) or CPA (Certified Public Accountant) which could be obtained by Individuals after passing standardized tests. Another innovative solution for the assessment and credentials that crossed countries boundaries is the digital badges which introduce much entertainment for online educational activities and experiences.

B. Match Making

Based on their survey that covered more than 100 initiatives in 25 countries, Mourshed et al. [11] observed that there are many cases that educational providers have built strong relationships with employers so that they can hire their graduates immediately after graduation based on the matchmaking and recommendation process that is being done by the educational providers themselves. With current technological advancement, matchmaking could be a game changer in the employment scene.

Flanagan [16] also agreed that Tinder-style matchmaking is beneficial in the job market as well and shed the light on a similar app called "Emjoyment" which allow job seekers to swipe job posts which includes major highlights about the company, location and only one sentence job description and once the job seeker find a good post, he just hits "like". On the other side, employers start to see job seekers who liked their opportunity in a form of cards including resumes main highlights and if the recruiter found an interesting profile, he also hits "like" andat that moment both parties connect together at a push of a button. This kind of matchmaking innovations could decrease the time lost in job applications and finding a good candidate and create direct engagement between employers and job seekers.

4. ARTIFICIAL INTELLIGENCE SYSTEM

The conceptual blueprint using artificial intelligence system include several components which can be integrated as one complete artificial intelligence system [17]. These are the standard components [18]:-

- Reasoning − It is the set of processes that empowers us to provide basis for judgement, making decisions, and prediction.
- Learning − It is the activity of gaining information or skill by studying, practising, being educated, or experiencing something. Learning improves the awareness of the subjects of the study.
- 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.
- Perception − It is the way of acquiring, interpreting, selecting, and organizing sensory information.
- 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.

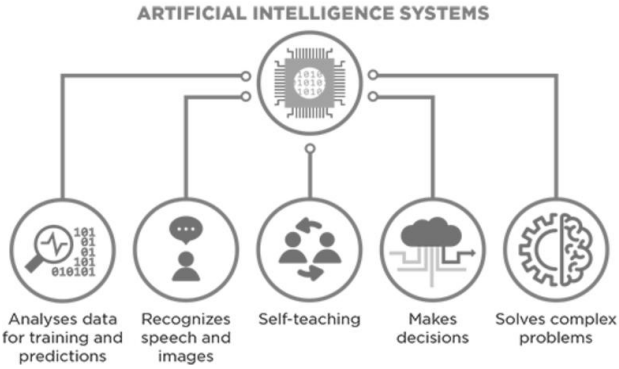


Figure 5. The artificial intelligence online learning system components.

5. CONCLUSIONS

This research was proposed by understanding the inter-relation between education and employment, what blueprints and systems that worked, and how AI can impact in the education for employment process whether by introducing new innovations that foster students learning process and placement in the job market or by harming the process and introducing unintentional bias, privacy breach, escalating power consumption and replacing human in workplaces. This paper is assessing new blueprints for learning and employability and how AI can fit in and foster the process, so further studies and experiments should be carried out to ensure the effectiveness of the blueprint and the accuracy of results of the AI application on the learning and employability process.

REFERENCES

[1] World Commission on Environment and Development (1987). *Our Common Future.* Oslo.

[2] Tuomi, I. (2018). *The Impact of Artificial Intelligence on Learning, Teaching, and Education.* Publications Office of the European Union.

[3] Popenici, S., & Kerr, S. (2017). Exploring the impact of artificial intelligence onteaching and learning in higher education. *Research and Practice in Technology Enhanced Learning.*

[4] Bayne, S. (2015). Teacherbot: interventions in automated teaching. *Teaching in Higher Education*, 455-467.

[5] Schmidt, A. (2017). *How AI Impacts Education.* Retrieved February 2019, from Forbes:https://www.forbes.com/sites/theyec/2017/12/27/how-ai-impacts- education/#22edd83f792e.

[6] Hawksworth, J. (2018). *AI and robots could create as many jobs as they displace.* Retrieved 2019, from World Economic Forum: https://www.weforum.org/agenda/2018/09/ai-and-robots-could-create-as-many-jobs-as-they- displace/.

[7] Andrews, W. (2018). *Craft an Artificial Intelligence Strategy: A Gartner Trend Insight Report.* Gartner, Inc.

[8] Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., & Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies* , 6(4), 100.
[9] *The Future of Jobs* (2018). Centre for the New Economy and Society.
[10] Martens, B., & Tolan, S. (2018). *Will this time be different? A review of the literature on theImpact of Artificial Intelligence on Employment, Incomes and Growth*. Digital Economy Working Paper 2018-08; JRC Technical Reports.
[11] Mourshed, M., Farrell, D., & Barton, D. (2019). *Education to employment: Designing a systemthat works*. Retrieved April 1, 2019, from McKinsey Center for Government: https://www.mckinsey.com/industries/social-sector/our-insights/education-to-employment-designing-a-system-that-works.
[12] Boyer, R. H., Peterson, N. D., Arora, P., & Caldwell, K., (2016). Five Approaches to Social Sustainability and an Integrated Way Forward. *Sustainability, 8(9), MDPI AG*.
[13] Brunello, G., & Rocco, L. (2017). The effects of vocational education on adult skills, employment and wages: What can we learn from PIAAC? *Springer Link*, 8-315.
[14] Schultz, T. W. (1961). Investment in Human Capital. *The American Economic Review* , 1-17.
[15] Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy* , 9-49.
[16] Flanagan, J. (2014). *Tinder-style matchmaking helps you bag your next job*. Retrieved from New Scientists: https://www.newscientist.com/article/dn25172-tinder-style-matchmaking- helps-you-bag-your-next-job/.
[17] Gus Bekdash, (2019). Using Human History, Psychology, and Biology to Make AI Safe for Humans, Chapman & Hall/CRC.
[18] 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, Hewlett-Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.

AUTHOR INDEX