Center for Artificial Intelligence Technology
Universiti Kebangsaan Malaysia

Universiti Kebangsaan Malaysia
The National University of Malaysia

FAKULTI TEKNOLOGI & SAINS MAKLUMAT
Faculty of Information Science & Technology

VIRTUAL CONFERENCE

MCAIT 2021

4 - 5 AUGUST

The 5th International Multi-Conference on Artificial Intelligence Technology

Artificial Intelligence in the 4th Industrial Revolution

e-PROCEEDINGS

Organized by Center For Artificial Intelligence Technology (CAIT)
Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, Malaysia

www.ftsm.ukm.my/mcait2021

mcait@ukm.edu.my
mcaitukm

Cetakan Pertama / *First Printing*, 2021

Diterbitkan di Malaysia oleh/ *Published in Malaysia by*
Pusat Kajian Teknologi Kecerdasan Buatan (CAIT)
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan, MALAYSIA
http://www.ftsm.ukm.my/cait/
e-mel: cait.ftsm@ukm.edu.my

Perpustakaan Negara Malaysia

Data Pengkatalogan-dalam-Penerbitan /
Cataloguing-in-Publication Data

E-PROCEEDINGS OF THE 5TH INTERNATIONAL MULTI-CONFERENCE ON ARTIFICIAL INTELLIGENCE TECHNOLOGY (MCAIT2021) / Editors: Khairudin Omar, Azuraliza Abu Bakar, Syaimak Abdul Shukor, Bahari Idrus, Sabrina Tiun, Nazatul Aini Abdul Majid, Shidrokh Goudarzi, Mohd Syazwan Baharuddin.

eISBN 978-967-19332-1-3

E- Proceedings of the 5th International Multi-Conference on Artificial Intelligence Technology (MCAIT 2021)

Artificial Intelligence in the 4th Industrial Revolution

Editors

Khairudin Omar

Azuraliza Abu Bakar

Syaimak Abdul Shukor

Bahari Idrus

SabrinaTiun

Nazatul Aini Abdul Majid

Shidrokh Goudarzi

Mohd Syazwan Baharuddin

Center For Artificial Intelligence Technology (CAIT)
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia

FOREWORD FROM THE DEAN

Assalamualaikum and peace be upon all

We are delighted to conduct M-CAIT 2021, the 5th International Multi-Conference on Artificial Intelligence Technology. The M-CAIT 2021 conference is held biennially by the Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM). It is indeed a remarkable moment for the faculty because it marks the 5th series of this conference. The year 2021 has been challenging for everyone due to the COVID-19 pandemic. In adapting to these challenges, M-CAIT 2021 is the first virtual conference organized by our faculty members for an international audience.

The conference aims to obtain and extend the knowledge of the recent issues, opinions, and bright ideas about developing a comprehensive Artificial Intelligence (AI) technology. MCAIT 2021 invites scholars and encourages researchers to submit high-quality papers to this conference. It is here that we extensively share and exchange ideas, thoughts and discussions on all aspects of artificial intelligence, information technology, information systems, big data, emerging ICT applications, data analytics and cyber technologies. We hope that the collaborative atmosphere of this conference will facilitate the advancement of AI technology in the 4th Industrial Revolution among participants of the conference for improving the quality and benefits of their research.

It is a great pleasure to welcome all the participants of this virtual conference in UKM, Bangi. I hope that this conference will be a valuable forum for academicians, industry and scientists to share their precious research. This event will give significant contributions to the development of AI Technology and raise the awareness of scientific community members in bringing a better life.

I hope that the conference will be stimulating and memorable for you. Therefore, I would like to thank all the participants, presenters and committee members for their continued support for MCAIT-2021.

Thank you, wassalamu alaikum.

Prof Dr Salwani Abdullah
Dean
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

WELCOME MESSAGE FROM THE MCAIT 2021 CONFERENCE CHAIR

On behalf of the MCAIT 2021 and AWIST2021 organizing committees, I am delighted to welcome all participants of the 5th International Multi-Conference on Artificial Intelligence Technology (MCAIT 2021) and the 4th ASEAN Workshop on Information Science and Technology 2021 (AWIST2021) that take place virtually on August 4 and 5, 2021. MCAIT 2021 is being hosted in conjunction with AWIST2021 to invite more scholars to submit their research ideas and results. This event is organized by the Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science & Technology (FTSM), Universiti Kebangsaan Malaysia.

This conference will focus on artificial intelligence (AI) and related disciplines. Attendees will be able to share ideas, learn about new prospects, reconnect with past networks, and broaden their knowledge in this setting. The session will be held remotely due to the pandemic Covid19. The theme of MCAIT 2021 is "Artificial Intelligence in the 4th Industrial Revolution", which will cover a wide range of the topic such as artificial intelligence, semantic technology, ICT solution to the industrial problem, computer vision, machine learning, and others.

The MCAIT 2021 was established to encourage international scientific exchanges and collaborations among AI researchers. MCAIT 2021 is the fifth edition of MCAIT, which began in Putrajaya in 2011, and was followed by MCAIT 2013 in Shah Alam, M-CAIT 2016 in Malacca, and MCAIT 2018 in Sarawak, Malaysia. Meanwhile, the FTSM is hosting AWIST2021, a workshop co-hosted by the Japan Advanced Institute of Science and Technology (JAIST), Universitas Komputer Indonesia (UNIKOM), and Universiti Teknologi MARA (UiTM). The connection began in 2013 and has since been strengthened through a variety of activities. As a result, UKM values the submission of articles and contributions by AWIST collaborators.

The programme committee has approved 50 manuscripts for oral presentations based on reviewers' recommendations. All accepted articles will be published as an e-proceeding with ISBN registration and will be presented at MCAIT 2021. The authors of selected articles will be invited to submit an extended work for publication in the "Asia-Pacific Journal of Information Technology and Multimedia (APJITM)," which is indexed by DOAJ, MYCITE, and others (see https://www.ukm.my/apjitm/indexing).

During the conference, we will hear from six keynote speakers who will offer their wealth of knowledge and expertise. We're delighted to have:

- Professor Dr Angelo Cangelosi of the University of Manchester in the United Kingdom, UK.
- Professor Dr Eleni Vasilaki of the University of Sheffield, UK.
- Professor Dr Rose Alinda Alias, from the University Teknologi Malaysia.
- Dr Bassam Al-Salemi, a Senior Data Scientists from the PETRONAS.
- Professor Dr Azuraliza Abu Bakar from the Universiti Kebangsaan Malaysia.
- Dr Mohd Ridzwan Yaakub from the Universiti Kebangsaan Malaysia.

I'd like to express my gratitude to all of the authors who submitted papers to MCAIT 2021, as well as the members of the programme committee and reviewers, for their time and work in reviewing the manuscripts. We also appreciate AWIST 2021's participation in MCAIT 2021. Without their efforts, the MCAIT 2021 would not have been possible. Finally, we'd like to thank the British High Commission in Kuala Lumpur for sponsoring our UK keynote speakers.

Professor Dr Masri Ayob
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

PREFACE

The 5th International Multi-Conference on Artificial Intelligence Technology (MCAIT2021) is organized by CAIT, a center of excellence of the Faculty of Information Science and Technology, UKM offers a collaborative environment to academicians, researchers and practitioners to exchange and share their experiences in the advancement of AI technology.

This proceeding contains papers presented at the MCAIT2021 conference held virtually from August 4 - 5, 2021, with the theme '*Artificial Intelligence in the 4th Industrial Revolution'*. The MCAIT2021 conference addresses a broad range of research and practical topics, including artificial intelligence, semantic technology, ICT solution to industrial problems, computer vision, machine learning, etc. Each contributed paper was refereed before being accepted for publication in these proceedings. The papers were accepted for publication based on their originality, significance, technical content and application contents on the conference's topics.

For this year, the MCAIT2021 features keynote talks of six prominent researchers with the scope of talks on; development robotics for language learning and trust, information system management, AI Roadmap in Malaysia, sparse reservoir computing, and data science. Around 50 papers were presented in six parallel sessions, with each session having six to ten regular paper presentations. The six parallel sessions were AI related to *Business, Education, Healthcare, Human-being, Intelligence in AI* and *Robots*. In addition, nearly 100 registrants of the conference came from across the country and represented government, industry, and academia.

On behalf of the MCAIT2021 organizer, we would like to take this opportunity and warmly thank all the authors who submitted their work to MCAIT2021. Their contributions to the conference are greatly appreciated, and we are looking forward to their continued contributions to future MCAIT conferences.

Editorial Committees
MCAIT2021
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

Table of Contents

THEME: EDUCATION

THEME: HEALTHCARE

THEME: HUMAN WELL-BEING

THEME: ROBOT

The Role of IT and IS in Improving Humanitarian Supply Chain Management

Prima Denny Sentia[a,b], Syaimak Abdul Shukor [a]*, Amelia Natasya Abdul Wahab[a], Muriati Mukhtar[a]

[a]Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia
[b]Industrial Engineering Department, Universitas Syiah Kuala, Jl. Tgk. Abdur Rauf No. 7 Darussalam, Banda Aceh 23111, Indonesia
*Email: syaimak@ukm.edu.my

Abstract

Natural disasters that occur worldwide attract researchers to find fast and targeted ways to distribute aid. Since 2007, studies on Information Technology (IT) and Information Systems (IS) have been growing and advancing to assist Humanitarian Supply Chain Management (HSCM). By using qualitative methods based on the literature review, this article has two objectives. The first is to explain IT and IS's role in improving the humanitarian logistics supply chain (HSCM). Second, exploring current trends in the use of IT and IS in several humanitarian organizations. The results of this study indicate that IT and IS have at least five prominent roles, namely producing transparent HSCM, minimizing risk, reducing costs, detecting demand for aids, and helping stakeholders to make the right policies. Therefore, humanitarian organizations are encouraged to take advantage of IT and IS in carrying out their humanitarian activities.

Keywords: Information Technology; Information System; Humanitarian Supply Chain Management.

1. Introduction

Every year, various aid worth hundreds of millions of dollars is sent to affected countries worldwide (PSA, 2015). There are 83% of disasters that have been caused by natural hazards that cause extreme weather, floods, storms, landslides, and heatwaves. This situation occurred worldwide, killed more than 410,000 people, and affected 1.7 billion people (IFRC, 2020). Various kinds of natural disasters that have occurred worldwide and have increased over the last two decades have become a concern for researchers concerned in the field of the humanitarian supply chain (HSC) (Upadhyay *et al.*, 2020).

During a disaster, HSCM plays an essential role in allocating assistance to disaster victims, especially in the logistics sector. The supply chain must be operated efficiently to ensure the logistics distribution reaches the right target. Response to disasters involves government, non-governmental organizations (NGOs), UN agencies, the military, and private sector organizations (Costa*etal.*,2012). Benini*et al.*(2009), in their research, states that there are two pieces of information needed to distribute humanitarian logistics: First, it needs information which covers needs assessment, size of the affected population and additional adversity vulnerability, damage levels, pre-existing poverty. Second, logistics information (distance from the hub, transportation capacity on a given day, access roads open on a given day, and existing cargo handling orders. During a disaster situation, decision-makers need a precise picture of the interconnection between systems. In this case, HSCM often experiences delays in decision making, resulting in delays in rescue teams in distributing logistics to victims (Tatham and Houghton, 2011).

Awareness of the difficulties in logistics distribution to disaster victims led researchers to focus their studies on the HSCM study. The research has implications for the increasing number of publications in the HSCM field, and they are studied in various sub-topics. One of them is related to information technology and information

systems (IT / IS) to improve the humanitarian supply chain. This article will explain IT and IS's importance for HSCM and see IT and IS trends in the HSCM process stages.

2. Recent Studies

Studies on information technology (IT) and information systems (IS) in the humanitarian supply chain have started to increase since 2007 (Behl and Dutta, 2019). Several researchers have focused their studies on IT use and are to improve the humanitarian logistics supply chain (Kauremaa *et al.*, 2004; Howden, 2009; Overstreet *et al.*, 2011; Chingono and Mbohwa, 2016; Comes and Walle, 2016). IT components include hardware, software, electronics, the internet, and other telecommunication equipment, enabling storage, retrieval, transmission, and data manipulation. Meanwhile, IS studies how to build and integrate information technology solutions with existing businesses to achieve goals effectively and efficiently. Both It and IS are the main drivers in logistics so that the logistic system can be automated.

In general, the use of IS and IT in the humanitarian supply chain will improve the flow of information that can efficiently integrate logistic and non-logistics units and provide reasonable bail to donors. In this context, information systems can improve humanitarian operations' continuity by sharing information through collaboration between organizations. Information systems in the humanitarian supply chain also can reduce corruption and market distortions during humanitarian operations (Howden, 2009).

The study of Comes and Walle (2016) explains that the purpose of information systems in humanitarian aid is to assist humanitarian logisticians in reaching logistical decisions that must be taken. It means that the information system must be designed to collect information in a timely, correct, and relevant manner about several things: first, disaster information (concerning human needs, gaps, and priorities, resources, funding, availability of access, and infrastructure). Second, the supply chain (relating to capacity, the status of goods and delivery, actors in the supply chain, suppliers, transportation and infrastructure providers, and end-users). In disaster management, the use of a humanitarian information system has three implications: (1) As an assessment tool to assess the general description of human needs and the field's situation, and the resources required. In the assessment context, information systems will need to support community resilience; (2) Coordination among all humanitarian activities to avoid overlaps; (3) Humanitarian briefs and appeals towards donors.

Overstreet et al. (2011) conducted other research using the theory of constraints and the IS literature management to study logistics distribution in a disaster. This study builds a research framework using logistics elements: organization's personnel, infrastructure, planning/policies/procedures, transportation, information technology, and inventory management in humanitarian aid. Through the above research, it can be concluded that IT and IS are essential components so that the logistic system can update data, plan, and estimate information throughout the supply chain so that the desired results are maximized (Lindenberg and Bryant, 2001; Mubaraka*et al.*, 2013; Kabra and Ramesh, 2015).

3. Information Technology and Information Systems: Why Important?

IT and IS's role in HSCM management is important in data availability as an essential basis for determining policies when a disaster occurs. The ability of IT / IS to provide relevant information will result in an appropriate analysis in optimal supply chain operations. Some of the critical roles of IT / IS in HSCM are outlined as follows:

a. Using IT and IS can produce a transparent supply chain. IT and IS will continuously increase productivity—all data, information, and how the supply chain works can be seen by all agents involved. Transparency of information will facilitate the planning, coordination, and collaboration processes between agents.
b. IT and IS can provide technology to optimize all possible scenarios in the supply chain. Therefore, supplies will be more integrated and coordinated and able to minimize risks.
c. The emphasis on costs can be done with IT and IS's help to analyze trends that may recur and may emerge

in the future. Thanks to this analysis, the optimal use of resources and assets reduces costs.

d. Detection of the demand for needed aids will be easier if it uses IT and IS. Stakeholders can provide targeted assistance according to the needs of disaster victims. The utilization of IT and IS can be used to conduct early awareness analysis. This analysis helps what type of relief items the victim needs most and how much is required. Estimating requests like this will make it easier for each agent to assist victims.
e. IT and IS will be a source for making the right decisions. The utilization of IT will help collect data, transcribe complex data into simpler forms, and make it a report to support decision making. Based on these roles, it can be said that IT and IS are the most critical determinants of the success of HSCM.

4. Current Trends

At present, the adoption of IT and IS in HSCM seems to be a universal imperative so that aid organizations can be handled effectively and efficiently(Avgerou, 2001). Several humanitarian organizations and institutions dealing with disaster relief have used IT and IS to automate their supply chains. The World Food Program (WFP) is a food aid organization of the United Nations. WFP has implemented an effective response system for several major disasters, such as the Iran earthquake (the 1960s). They also helped distribute food for Elnino victims in Southern Africa in early 1990. Assist people affected by drought in Horn Africa and Ethiopia in 2000 WFP is one organization that uses IS to carry out its activities. WFP initially did not have comprehensive software for tracking and tracing its commodities' status, movements, and location. However, since February 1996, the WFP executive board built a new Corporate Supply Chain Tracking System (COMPAS) and officially implemented it in January 1997. This system is used to develop, implement, record, monitor, and report the entire supply chain process. This system also allows for comprehensive, accurate, and timely computation and reporting of commodities. The benefits of using COMPAS can be felt by donors, the government, and society. This system can monitor many things, such as information about donations, shipping, vessel discharges, stock stored, stock transit, loss, and distribution (Tusiime and Byrne, 2011).

Save the Children organization has chosen the Helios system, which the Fritz Institute sponsors. Also, they built their in-house system with simple additions to Agresso, a commercial ERP system used in the financial sector. Care USA uses software designed by Aid Matrix, which is known to be active in developing software for the Aids industry. All aims to optimize the process of distribution of aid to disaster victims(PSA, 2015).

Oxfam is a British non-profit organization that focuses on the alleviation of global poverty. Oxfam also collaborates with communities before, during, and after crises to build their resilience, save lives, and address the root causes of conflict and disaster. Oxfam utilizes some technologies and software in carrying out its activities. Oxfam used the Scaling Humanitarian ICTs Network (SHINE) to scale information communication technologies to improve their humanitarian programs' quality and efficiency. SHINE is mainly used to increase the effectiveness of humanitarian delivery. SHINE is capable of performing accurate assessments through mobile data and collection tools.

In July 2016, Oxfam launched the mobile survey toolkit named Mobenzi as a data collection tool. Besides, Oxfam has also used World Visions Last Mile Mobile Solutions (LMMS). This technology was developed by Last Mile Mobile Solutions, which is designed to combine software and hardware applications. The purpose of using this tool is to simplify beneficiary registration, verification, planning, distribution management, monitoring, and reporting (O'Donnell, 2017). Humanitarian organizations must adapt to a dynamic and complex humanitarian environment and coordinate with all agents in HSCM(L'Hermitte *et al.*, 2016; Sigala*et al.,* 2019). Then,IT and IS's use becomes a fundamental solution for managing information flow in disaster management. With IT and IS's help, humanitarian organizations can carry out their activities quickly and on target.

5. Conclusion

The arrival of a disaster is an unpredictable situation. Meanwhile, handling victims, especially in aid distribution, must be carried out as quickly as possible. Thus, previous researchers focused their studies on the use of IT and IS in improving humanitarian responses. Several humanitarian organizations adopted the use of IT and IS to produce an integrated HSCM. With the help of IT and IS, the transparency of the supply chain will be maintained. Information at each level will be easily monitored, thus facilitating coordination between agents. IT and IS can generate scenarios so that agents can analyze possible risks in the supply chain, analyze trends, and request assistance. Most importantly, IT and IS will assist stakeholders in making policies, especially those related to disaster assistance distribution.

References

Avgerou, C. (2001). The significance of context in information systems and organizational change. Information Systems Journal, 11(1), pp. 43–63.

Behl, A. and Dutta, P. (2019). Humanitarian supply chain management: a thematic literature review and future directions of research. Annals of Operations Research. Springer US, 283(1–2), pp. 1001–1044.

Benini, A., Conley,C., Dittemore, B. and Waksman, Z. (2009). Survivor needs or logistical convenience? Factors shaping decisions to deliver relief to earthquake-affected communities, Pakistan 2005-06. Disasters, 33(1), pp. 110–131.

Chingono, T. and Mbohwa, C. (2016). Information technologies for humanitarian logistics and supply chain management in zimbabwe. Proceedings of the International Conference on Industrial Engineering and Operations Management, pp. 1038–1046.

Comes, T. and Walle, B. Van De (2016). Information Systems for Humanitarian Logistics. Supply Chain Management for Humanitarians: Tools for Practice. United Kingdom: Kogan Page, pp. 257–284.

Costa, S. R. A. da, Campos, V. B. G. and Bandeira, R. A. de M. (2012). Supply Chains in Humanitarian Operations: Cases and Analysis. Procedia - Social and Behavioral Sciences, 54, pp. 598–607.

Howden, M. (2009). How Humanitarian Logistics Information Systems Can Improve Humanitarian Supply Chain: A View from the Field.Proceedings of the 6th International ISCRAM Conference, pp. 90–91.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2020). Tackling the humanitarian impacts of the climate crisis together, World Disaster Report 2020.

Kabra, G. and Ramesh, A. (2015). Analyzing ICT issues in humanitarian supply chain management: A SAP-LAP linkages framework. Global Journal of Flexible Systems Management, 16(2), pp. 157–171.

Kauremaa, J., Auramo, J., Tanskanen, K. and Kärkkäinen, M. (2004). The use of information technology in supply chains : transactions and information sharing perspective. LRN Conference. Dublin, pp. 1–8.

L'Hermitte, C., Tatham, P., Brooks, B. and Bowles, M. (2016) 'Supply chain agility in humanitarian protracted operations. Journal of Humanitarian Logistics and Supply Chain Management, 6(2), pp. 173–201.

Lindenberg, M. and Bryant, C. (2001). Going Global. Transforming Relief and Development NGOs', Kumarian press, 295, p. 271.

Mubaraka, C. M., Kalulu, R. A. and Salisu, M. J. (2013). Information Technology and Humanitarian Emergency Response Management in Wfp Uganda : a Behavioral Perspective', 2(3), pp. 147–153.

O'Donnell, A. (2017). ICTs in Humanitarian Response: A learning review of a three-year, five-country programme. Available at: www.oxfam.org.

Overstreet, R. E., Hall,D., Hanna,J. B. and Kelly-Rainer,R. (2011). Research in humanitarian logistics', Journal of Humanitarian Logistics and Supply Chain Management, 1(2), pp. 114–131.

PSA, W. P. (2015) Humanitarian supply chain information systems : insights for successful implementation March 2015. Available at: https://www.pamsteele.co.uk/wp-content/uploads/2018/08/PSA-Humlog-IT-WP.pdf.

Sigala, I. F., Kettinger, W. J. and Wakolbinger, T. (2019). Digitizing the field: designing ERP systems for Triple-A humanitarian supply chains. Journal of Humanitarian Logistics and Supply Chain Management, 10(2), pp. 231–260.
Tatham, P. and Houghton, L. (2011). The wicked problem of humanitarian logistics and disaster relief aid', Journal of Humanitarian Logistics and Supply Chain Management, 1(1), pp. 15–31.
Tusiime, E. and Byrne, E. (2011). Information Systems Innovation in the Humanitarian Sector. Information Technologies & International Development, 7(4), pp. 35-52.
Upadhyay, A., Mukhuty,S., Kumari, S., Garza-Reyes,J. A. and Shukla, V. (2020) 'A review of lean and agile management in humanitarian supply chains: analysing the pre-disaster and post-disaster phases and future directions. Production Planning and Control, pp. 1–14.

RFID-based Payment System Using e-KTP in Shop with Membership System

Syahrul[a*], Mochamad Fajar Wicaksono[b], M. Rinaldi Hasanudin[c]

[a,b,c]Program Studi Sistem Komputer, Universitas Komputer Indonesia, Jl. Dipati Ukur 112-116 Bandung, Indonesia
*Email: syahrul@email.unikom.ac.id

Abstract

This paper describes a design that applies the e-money system through the use of e-identity cards (hereinafter referred to as e-KTP). The aim is not only to make transactions easier and to make transaction time efficient, it also reduces the use of cash as a form of security in carrying large amounts of money and being inconvenienced with refunds in small denominations. The method used in this research is design and implementation. This system is made by utilizing e-KTP based on Radio Frequency Identification and equipped with additional security in the form of Fingerprint Biometrics. From the results of testing several e-KTP samples, all of them were successfully recorded in the database and successfully read again. Data that has been successfully recorded into the database is in the form of user identity, nominal money deposited and the balance after the transaction. This system is expected to be integrated into the integrated and multifunctional e-KTP system so that we are no longer bothered with having to carry a variety of other cards.

Keywords: e-KTP; RFID; e-payment; member system

1. Introduction

Transactions are activities carried out by a person that can result in the exchange of goods or money, both increasing and decreasing. However, the large number of exchanges using paper and metal money makes transactions inefficient because we have to carry a lot of money, this makes money pile up in the wallet and uncomfortable when using it. A lot of resources are wasted because money is damaged or obsolete due to age or improper use. To reduce the use of money and save manufacturing resources, the developer is currently implementing Radio Frequency Identification technology which aims to make transactions easier by simply utilizing the e-KTP (electronic identity card) provided by the government.

An automation system that controls the supply of goods for supermarkets in Nigeria using RFID which operates at a frequency of 13.56 MHz (Boyinbode & Akinyede, 2015). Likewise, an RFID-based smart trolley system has been developed which can provide a display on the trolley in the form of the amount of money to be paid based on the number of items placed in the cart. The goal is to save queuing time so that when you arrive at the cashier, you just have to pay immediately without the need to scan goods via bar codes. (Devi, Kaarthik, Selvi, Nandhini, &Priya, 2017). In a paper written by Gunasagar et al, they describe a smart billing system that utilizes RFID and weight sensors. In this system, the function of RFID is to check each product placed in the trolley and display the bill amount on the LCD (Gunasagar, &Balachander, 2020). A smart trolley system that is used for shopping in malls and shopping centers was developed (Machhirke, Priyanka, Rathod, Petkar, &Golait, 2017).

In these papers, a system that uses e-KTP is proposed for a home security system, namely by using the RFID technology embedded in the e-KTP to be able to open and close house doors (Andriansyah et al., 2017), Mustolih, Lenggana, & Mulyana (2019), Putra, Marwanto, & Qomaruddin (2017). A system that the use of the e-KTP system should apply a decentralized network model, especially for storing e-KTP data without violating privacy laws (Awangga, Harani, &Setyawan, 2019). In the research conducted by Fajri, et al, they proposed an

integrated posyandu application with android-based communication technology. The use of RFID allows e-KTP with the data collection process to be done quickly and accurately using an android smartphone (Fajri, & Oktaviana, 2019). Meanwhile, Sari et al. (2018) proposed the use of e-KTP to conduct e-voting in general elections using electronic media and the internet which aims to facilitate and speed up the selection process by adding fingerprint identification as data validation (Sari et al., 2018).

However, in our study, here is developed a system which functions to reduce the use of money and save resources. Its manufacture applies Radio Frequency Identification technology which aims to make transactions easier by only using the electronic Identity Card (e-KTP) that has been provided by the government. This system implements the use of e-KTP to replace the use of banknotes and coins which make transactions inefficient because we have to carry a lot of money, causing inconvenience in transactions. This system is used specifically for transactions in the store on a subscription basis as a member.

2. Methods

The method used in this research is design and implementation. The initial stage of making this system is by designing in determining system requirements. At this stage, what needs can be prepared that must be met in building a system that implements e-KTP as the basis of RFID. Figure 1 shows a block diagram of the system being built.

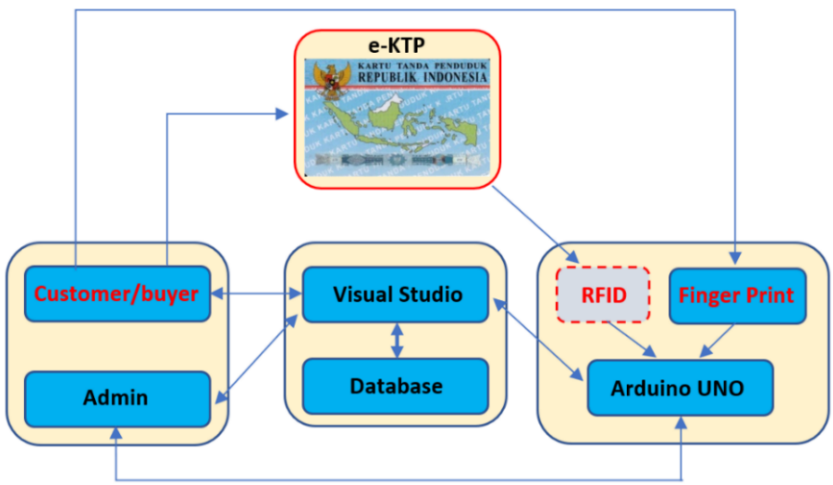


Fig. 1. System block diagram

The following is an explanation of the function of each sub-block of the system block diagram: Arduino UNO functions to process data read from sensors, RFID reader is used to scan E-KTP, finger print is used to scan fingerprints, E-KTP is used as RFID Tag, customer or buyer is the person who will make the transaction, admin is in charge of managing all systems and filling funds, Visual Studio serves as a display interface to buyers or admins, and the database serves to store data from buyers and stored funds.

The system works with two sensors, namely Fingerprint and RFID reader which will then be processed by Arduino UNO. When the user wants to use this system, the user must register himself in the form of an e-KTP and fingerprint to the admin then complete the registration column. Once registered, the user can fill in a virtual balance by exchanging real money to the admin according to the amount. When the virtual balance has been filled, users can make transactions where the system will deduct the balance according to the price of the goods. Withdrawal of balances must attach e-KTP and fingerprints to the tools provided as a transaction requirement.

After the transaction is complete, the user can view the purchase history to the admin by submitting the e-KTP as user identity.

Electronic Identity Card (e-KTP) is a modern card made based on RFID (radio frequency identification). This card, both in terms of physical and usage, has a computerized function. Each e-KTP has a different unique code that will be read using an RFID reader and processed by a microcontroller. Furthermore, the data will be compared with data registered in the database. (Fajri, & Oktaviana, 2019). The e-KTP program in Indonesia began in 2009 with the designation of four cities as national pilot projects. The four cities are Padang, Makassar, Yogyakarta and Denpasar.

The use of fingerprint identification as data validation during financial transactions or when payments are made. AS608 Optical Fingerprint is a finger scanner. How AS608 works: finger is placed on the surface of the glass prism and light occurs through the face of the other prism. The angle of incidence is greater than the critical angle and therefore all light is actually reflected internally from the troughs of the finger. However, the ridge absorbed most of the light. In this way the valley appears bright and the ridge dark. This results in a high contrast fingerprint image.

3. Results and discussion

3.1. Hardware Testing

Tests carried out on a payment system using e-KTP based on Radio Frequency Identification apply Visual Basic.Net and Arduino testing, where tests were carried out on the application. The trick is to test whether the functions contained in the tool have functioned as desired. From the results of the tests carried out, all the sub blocks and the system work well. In Table 1, a list of Payment System Testing using e-KTP is shown.

Table 1. List of Payment System Testing Using e-KTP

Test item	Data tested
Login admin	Enter your ID and password
Registration	Registering user data
Payment	Change data and add balances
Purchase history	Reducing the balance according to the price of the goods

3.2. Testing on the Software Interface

Testing the software interface is done using a computer by opening Visual Basic.Net which functions as an interface from the system to the user. Furthermore, to be able to display payment results, customers or buyers must visit the cashier who manages the application. After successfully registering and filling the balance, the customers / buyers can make a payment. Figure 2 shows the payment interface.

The purpose of the Admin testing is to ensure that the admin function is running correctly when the admin updates data and adds user balances. Admin testing can be seen in Figure 3 as updating data and filling balances.

4. Conclusion

The resulting hardware system successfully reads and receives data input from the e-KTP and fingerprint validation. The system built successfully registers user data. The system was successfully built to increase and decrease the balance according to the simulation of the price of goods. The resulting system can store data, balances and transaction history of the user. The resulting system can show data and balances via admin, then

display transaction history. This system is expected to replace the use of real money, making transactions easier.

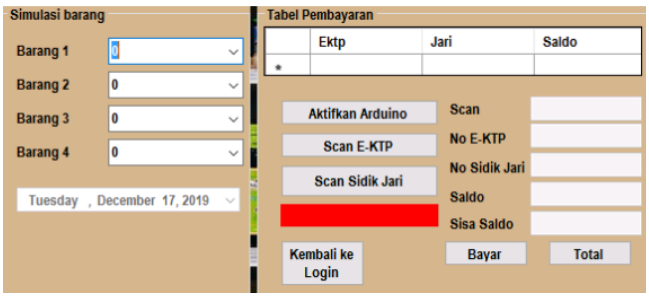


Fig. 2.Interface for payment

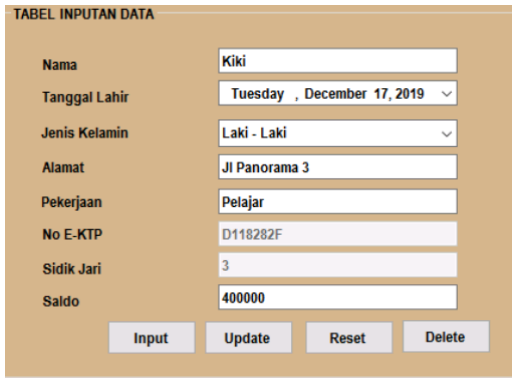


Figure 3.Interface for data updating and balance filling

References

Andriansyah, M., Subali M., Purwanto, I., Pramono, R. A., & Antonius I. S. (2017). e-KTP as the basis of home security system usingarduinouno. Retrieved from https://www.researchgate.net/profile/Miftah_Andriansyah/publication/323952948_e-KTP_as_the_basis_of_home_security_system_using_arduino_UNO/links/5b63cc7d0f7e9b00b2a24c01/e-KTP-as-the-basis-of-home-security-system-using-arduino-UNO.

Awangga, R. M., Harani, N. H., &Setyawan, M. Y. H. (2019). High interoperability e-KTP decentralised database network using distributed hash table, 17, 1360-1366

Boyinbode, O., &Akinyede, O. (2015). A RFID based inventory control system for nigerian supermarkets. international journal of computer applications, 116, 7-12.

Devi K. G., Kaarthik, T. A., Selvi N. K., Nandhini K., & Priya S. (2017). Smart shopping trolley using rfid based on IoT. International journal of innovative research in computer and communication engineering, 5, 5392-5398.

Fajri, I. A., & Oktaviana, S. (2019). NFC technology on integrated e-KTP for health service post in indonesia based on android. Jurnalmultinetics, 5, 114-119.

Gunasagar, T., &Balachander, B. (2020). Smart billing system using rfid and weight sensors. International journal of advanced research in engineering and technology, 11, 325-329.

Machhirke, K., Priyanka, G.,Rathod, R., Petkar, R., &Golait, M. (2017). A new technology of smart shopping cart using RFID and ZIGBEE. International journal on recent and innovation trends in computing and communication, 5, 256-259.

Mustolih, R., Lenggana, U. T., & Mulyana, J. (2019). Utilization of e-ktp as home safety using arduinonano based on android. JOIN (Jurnal Online Informatika), 4, 9-15.

Putra, Y. A., Marwanto, A., & Qomaruddin, M. (2017). Design of residental safety system using E-KTP. Journal of telematics and informatics (JTI), 5, 1-9.

Sari, R. D. Y., Sindung H.W.S., & Sri, A. K. (2018). Design and e-voting information system based web using e-ktp on election of the head of city of Semarang. Journal of applied information and communication technologies, 3, 1-6.

Web Based Information System Of Test of English as a Foreign Language (TOEFL)

Novrini Hasti[a]*, Sekar Rhiandari Graitasadu[b]

[a,b]Universitas Komputer Indonesia, Jl. Dipatiukur No. 112, Bandung, 40132, Indonesia
* Email: novrini.hasti@email.unikom.ac.id

Abstract

Currently the world has recognized a technology called the internet. With a global network, the internet can be accessed 24 hours a day. The world of education is also inseparable from this advancement in information technology. To improve education quality standards and graduate quality standards, TOEFL (Test of English as a Foreign Language) was held. The purpose of this research is to design and apply a web-based registration information system, scheduling, test execution and TOEFL test result validation processes. This is done so that the initial registration process, schedule selection, test implementation and the process of validating the test results become faster and easier. In this research, the systems approach method used is the object-oriented systems approach, with several tools and working techniques using UML which consists of use cases, activity diagrams, class diagrams, sequence diagrams and deployment diagrams. For the system development method using the Prototype method. The programming language used in designing and implementing the system is the CI programming language and the database used is MySQL.

Keywords: information system; object oriented method; prototype method

1. Introduction

One of the English language test models used to measure the ability of a person whose English is not used as native language is the Test of English as a Foreign Language (TOEFL) (Sharpe, 2007). Now, TOEFL is used in many countries as a university entry requirement. The need for TOEFL around the world is increasing (Nasir et al., 2019).

In this era of globalization, the quality of human resources is the key to competitiveness between countries. To take an active role in relations between countries globally, it is required to be involved in communication skills, especially in English. The logical consequence is that many foreign workers will visit and even look for work in Indonesia. In this case, Indonesia must immediately prepare human resources who are able to speak English actively (Shobikah, 2017).

To improve education quality standards, a study program in university really needs information that can facilitate university academic activities, especially issues of registration, scheduling and validation. Especially the process of the TOEFL. As one of the requirements, the TOEFL score will determine whether a student will graduate or not. To facilitate students in all TOEFL processes, universities must improve their services in this field.

There are several problems that occur in connection with the implementation of the TOEFL. To register for the TOEFL, students must come to the secretariat of the study program with a form and a photo. The form must be filled in regarding student registration data, test schedule and payment of the test. In addition, students must also come to campus to see the results of the test. To validate test results students must meet the coordinator of the study program. The test is carried out in an English literature study program that has not been computerized. The scoring system and score calculations are still done manually using a CAT machine to correct it. The things above resulted in ineffective and inefficient implementation of the TOEFL.

Previously there have been several studies related to the TOEFL. Related research about the TOEFL is the making of applications of English training (Marbun et al., 2016). Applications built based on web using the Waterfall method in the design process. A structured approach is used in this system design process. The features available in this system are the registration process and the process of implementing the TOEFL test which can be done on this system and accessed online and displays the test result scores. The difference with the research we do is, the use of the system development method, namely the prototype method, while the systems approach method uses the object-oriented method.

Another research related to the TOEFL information system is a study entitled "TOEFL Online Berbasis Web" (Rahman, 2016). This study aims to develop a web-based online TOEFL test system with the hope of supporting the implementation and implementation of the TOEFL test in accordance with the TOEFL test implementation standards. This research also used Waterfall method for system development method.

Apart from these studies, there are also studies that carry out the design and implementation of TOEFL and TPA applications with the waterfall method approach (Sugiri and Ramdhani, 2015). The purpose of this study is to produce an online-based application that can support the implementation of the TOEFL and TPA tests in universities. This research used Waterfall method for system development method and structured method for system approach method.

The purpose of this research is to design and apply a web-based registration information system, scheduling, test execution and toefl test result validation processes. This is done so that the initial registration process, schedule selection, test implementation and the process of validating the test results become faster and easier.

2. Research Methods

The research method is a scientific method or technique to obtain facts or principles from knowledge by collecting, recording and analyzing data based on science with specific purposes and uses. The system approach method used is Object-Oriented Method. The system development method used is Prototype Method. The stages of the prototype method are as below (Hasti et al., 2020):

1. Communication

   Making prototypes begins with communication between the software development team and customers, in this case the study program. The software development team will hold a meeting with the study program to determine the overall objectives for the software under development, identify specifications of what requirements already exist, and describe areas where system development is required.

2. Planning Quickly

   Prototyping iterations are planned quickly and modeling.

3. Fast design modeling

   Quick design focuses on representing all aspects of the software that will be seen by the end user.

4. Making of prototype

   The design will continue with the construction of a prototype.

5. Submission of systems / software to customers / users, shipping & feedback

   The prototype will be submitted to the study program, then they will evaluate the previously made prototypes, then they will provide feedback which will be used to perfect the specifications.

3. Results and Discussions

3.1 Overview of the proposed system

The proposed system is a development or improvement of the system that is currently running. Changes and improvements made such as changing and diverting business activities that have not used a computerized system

become computerized with the help of the website to be built. In addition, the new system adds activities or processes that can support the ongoing system in a better direction.

3.2 Proposed System Design

The design procedure of the proposed information system will be outlined in the following Use Case Diagram:

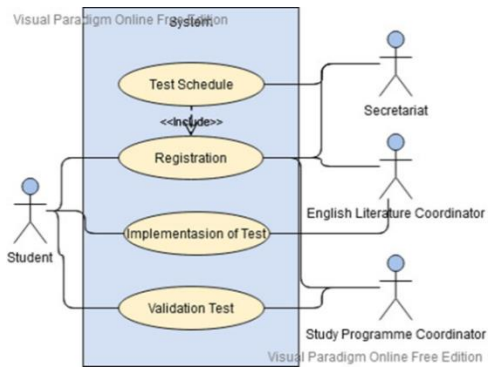


Fig 1. Usecase diagram proposed

The following table provides definitions of all actors and their descriptions.

Table 1. Definition of Actors and their descriptions

No	Actor	Description
1	Student	People who want to take the TOEFL test in English literature study programs for graduation requirements.
2	Secretariat	People who do toefl registration data collection. Liaising directly with the coordinator in the information system study program regarding test schedule info and details of costs and informing students via social media Facebook.
3	English Literature Coordinator	Orang yang mengkoordinir jalan nya tes *toefl* di sastra inggris. Mulai dari memberikan info tentang jadwal tes dan rincian biaya, menjadi pengawas pada saat tes berlangsung, mengkoreksi dan memberikan penilaian, serta membuat laporan hasil tes tiap gelombang untuk diserahkan ke koordinator prodi sistem informasi.
4	Study Programme Coordinator	The person who coordinates the course of his toefl in English literature. Starting from providing information about the test schedule and details of costs, being a supervisor during the test, correcting and giving an assessment, and making a report on the test results for each batch to be submitted to the study program coordinator.

3.3 Interface Design

Interface design is very important in making a program, because it is the basis for creating an interface that can provide convenience and not confusing for users in carrying out their activities. The interface design for the website menu structure for users and admins can be seen below:

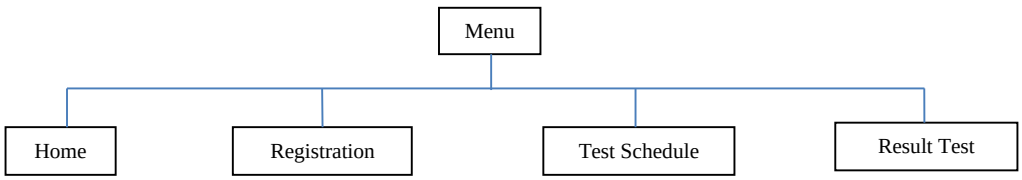


Fig 2. Website Menu Structure for Users

This menu structure contains the design of the menu interface for the user.

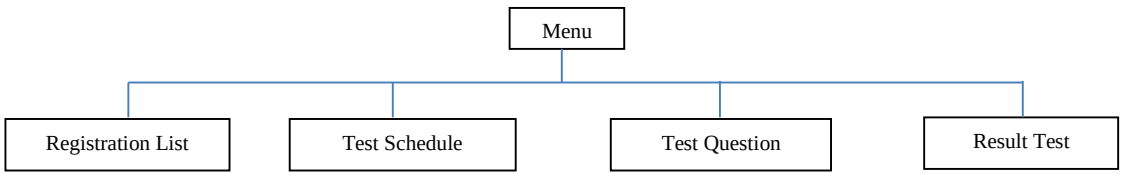


Fig 3. Website Menu Structure for Admin

This menu structure contains the design of the menu interface for the admin.

3.4 Interface Implementation

In a website, the implementation interface is designed in the form of a form with a php file extension. The following is a form designed based on the menu and sub menu :

Table 2. Implementation of the Student Main Page

Menu	Description	File Name
Home	Program files to display the main page	index.php
Registration	Program file to display registration page	daftar.php
Test Schedule	Program file to display test schedule page	jadwal.php
Test Result	Program file to display test result page	hasil.php

Table 3. Implementation of the Secretariat Main Page

Menu	Description	File Name
Login	Program files for handling logins	login.php
Student List	Program file to display student list page	daftarsiswa.php
Logout	Program files to exit the system	logout.php

Table 4. Implementation of the Admin Main Page

Menu	Description	File Name
Login	Program files for handling logins	login.php
Student List	Program file to display student list page	daftarsiswa.php
Score List	Program file to display score list page	daftarnilai.php
Question List	Program file to display question list page	daftarsoal.php
Question Code	Program file to display question code list page	jenissoal.php
Logout	Program files to exit the system	logout.php

4. Conclusions

By implementing this web-based TOEFL registration information system, it can help students, study program coordinator, study program secretariats and English literature coordinator in carrying out all TOEFL activities, such as registration, test schedules, test implementation and score validation. This application makes all activities run more effectively and efficiently.

Acknowledgment

The authors would like to thank everyone who helped in the implementation of this research.

References

Hasti, N., Dekiki, D., Gustiana, I., Wahyuni, W., Hartono, T. (2020). Web-Based Honorary Teacher Payroll Information System. Proceding The 3rd International Conferences on Informatics, Engineering, Science and Technology (Incitest). IOP Conference Series : Materials Science and Engineering. (879, 012021). IOP Publishing.

Marbun, Y. Y., Isnanto, R. R., & Martono, K. T. (2016). Pembuatan Aplikasi Toefl Sebagai Media Pelatihan Bahasa. Jurnal Teknologi dan Sistem Komputer. 4(1). 83–92.

Nasir, M., Hasyimi, H., Mahdi, M., Amru, A. (2019). Modelling System Test of English as a Foreign Language as a Web-Based Learning Media. International Conference on Science and Innovated Engineering (I-COSINE) IOP Conf. Series: Materials Science and Engineering. (536, 012130). IOP Publishing.

Rahman, M. G. (2016). Toefl Online Berbasis Web. Jurnal Teknika. 8(1), 781–790.

Sharpe, P. J. (2007). TOEFL Ibt, 12th edition. Barron's Educational, Inc

Shobikah, N. (2017). The Importance of English Language in Facing Asean Economic Community (AEC). At-Turats, 11(1). Retrieved from https://doi.org/10.24260/at-turats.v11i1.873.

Sugiri, U., & Ramdhani, M. A. (2015). Perancangan Dan Implementasi Aplikasi Toefl (Test Of English As Foreign Language) Dan Tpa (Tes Potensi Akademik) Berbasis Web Untuk Perguruan Tinggi. Jurnal Informasi. 7(1), 84–100.

Application of Web-Based Information System in Product Distribution

Lusi Melian[a*], Rifki Taufiqurrahman[b]

[a,b]UNIKOM, Jl. Dipati Ukur No. 112 - 116, Bandung, 40132, Indonesia
* Email: lusi.melian@email.unikom.ac.id

Abstract

PT. Bonli Cipta Sejahtera is a company engaged in large-scale pastry producers those quality has been guaranteed. The resulting products are distributed to distributors, agents and resellers throughout Indonesia. Constraints experienced in this company is the distribution system that has not been computerized so that it can slow down the distribution process. Companies need a distribution information system that can handle distribution problems in order to run smoothly. Interviews and observations are used to collect primary data, while documentation is used to collect secondary data. The research method uses a structured approach and the system development method uses a prototype model. The software used to build information systems, among others: Code Igniter Framework, Bootstrap Framework, XAMPP, Mozilla Firefox, Sublime Text Editor. The result of this research is an information system that regulates the distribution system to facilitate the work of the distribution division so that distribution can run more efficiently and data is well organized.

Keywords: Product Distribution, Information System, Prototype

1. Introduction

The technological developments provide considerable influence on an agency or company. Technology can be used for business processes that run in a company. Technology can provide the desired information from data that has been processed by information technology. The use of information technology has been widely used by many companies for transactions and reporting aimed at accelerating and simplifying all business activities. Information technology is a part of the information system that provides information to support the operation and management within an agency. The development of information technology helps to enhance and accelerate the accuracy of data into useful information (Purnamasari et al., 2014). The use of information systems in companies that have distribution divisions will help the company in every sales transaction, because every request for products that comes out will be recorded properly in the database system. With information systems, business processes in the company can run efficiently, effectively and well organized. The planning and implementation of information system are very important to business strategies on facilitate every activity (Damayanti, 2020).

PT. Bonli Cipta Sejahtera, located in Bandung, is one of the companies that produces large-scale pastries with guaranteed quality. The resulting products are distributed to distributors, agents and resellers throughout Indonesia. In the business process in this company, there is a distribution system for products through the distribution division. The products distributed by this company can reach thousands of units with destinations inside and outside the city per week. Making manual company passport is less effective and inefficient, because it takes a long time. The process of confirming the sent products is quite long because you have to wait for a return letter from the courier or an email confirmation from the package service. The duration of reporting is an obstacle faced by the distribution division because reporting still uses the manual method. Based on the above

problems, the company needs an information system that regulates the distribution system to run more efficiently and the data is well organized.

This study aims to build an information system with the ability to manage data and present the required information. The results of this study is an information system that is used to help provide solutions to problems that often arise in the process of distributing products. The proposed information system can solve the problem of making a company passport. The web-based information system in product distribution will automatically save the data distribution in the database. The products delivery confirmation process can be done by the consignee himself. The proposed information system can provide reports automatically. The availability of information system will improve the facilitate of administrative procedures, increase efficiency employee, improve output, and save time and money (Theorin et al., 2017).

2. Methods

In the design of information systems, it is necessary to use a methodology that can be used as a guideline for how and what to do during the manufacture of information systems. In this study, the systems approach method used is a structured approach method and the method of developing systems using the prototype development method.

A structured approach is a method of approach that is equipped with the tools and techniques needed in system development in order to obtain a structure based on a good and clear understanding (Ensour and Alinzi, 2014). The structured approach method is a method used to clearly define the system structure. Besides that, the structured approach method provides a clear picture of the data flow and describes the activities in detail (Kadir and Triwahyuni, 2013). The tools used in a structured approach include: flow maps, context diagrams, data flow diagrams (DFD), data dictionaries, and data design (normalization and table relations).

Systems development is preparing a new system in order to replace the old system or to revise the current system (Hartono, 2005). The system development method used is prototype method. Prototyping is an approach that creates a model that shows the features of a product, service, or proposed system (Hartono, 2005). Prototyping is a method in system development that can perform initial testing so that the system can be evaluated. Prototyping provides facilities for system developers to identify unmet needs and difficulties in using the system by interacting with each other between system developers and users during the manufacturing system.

3. Results and Discussion

a. Previous Research

Other studies are needed to complement current research in developing system as guidances and comparison. The author uses two previous studies including:

1. Susan Dian Purnamasari, Maulana, and Fatoni entitled Web Service-Based Products Distribution Information System. This research was conducted at PD Panca Motor Palembang, which is engaged in automotive dealerships, with branches spread across several regions. The previous process of distributing products was by sending products to several branches that were out of stock. In data processing, this company has used Enterprise Resource Planning (ERP) software. The application is not integrated with the branches, so it is not effective in exchanging information on products that only use the telephone. For this reason, a web service-based information system is built that can help and simplify the process of exchanging information flows (Sitanggan and Kusumaningrum, 2019).
2. Damayanti entitled Information Systems for the Distribution of Welding and Advertising Using the SCM Model. The research was conducted at Sahal Jaya Teknik, a company engaged in the manufacture of storefronts, canopies, ceilings, fences, and household furniture using aluminum and iron. The process of requesting goods from companies to suppliers still uses the telephone and the

process of recapping raw materials still uses notes. This is the cause of errors regarding the supply of raw materials. This research results an information system for the distribution of goods by applying the concept of Supply Chain Management, so that it can make it easier for companies in the process of requesting goods to suppliers, recording raw materials, and recording products to consumers (Sulthoni and Achilison, 2015).

b. Documents Analysis

The purpose of document analysis is to know and understand what documents are involved and flow in an ongoing system. Documents used in the information system of PT. Bonli Cipta Sejahtera are BJKB letter, invoice, company passport, package delivery letter and report.

c. Procedure Analysis

After collecting data and analyzing documents, then analyzing the procedure for products distribution using flow maps, context diagrams and data flow diagrams (DFD). As for the context diagram can be seen in Figure 1.

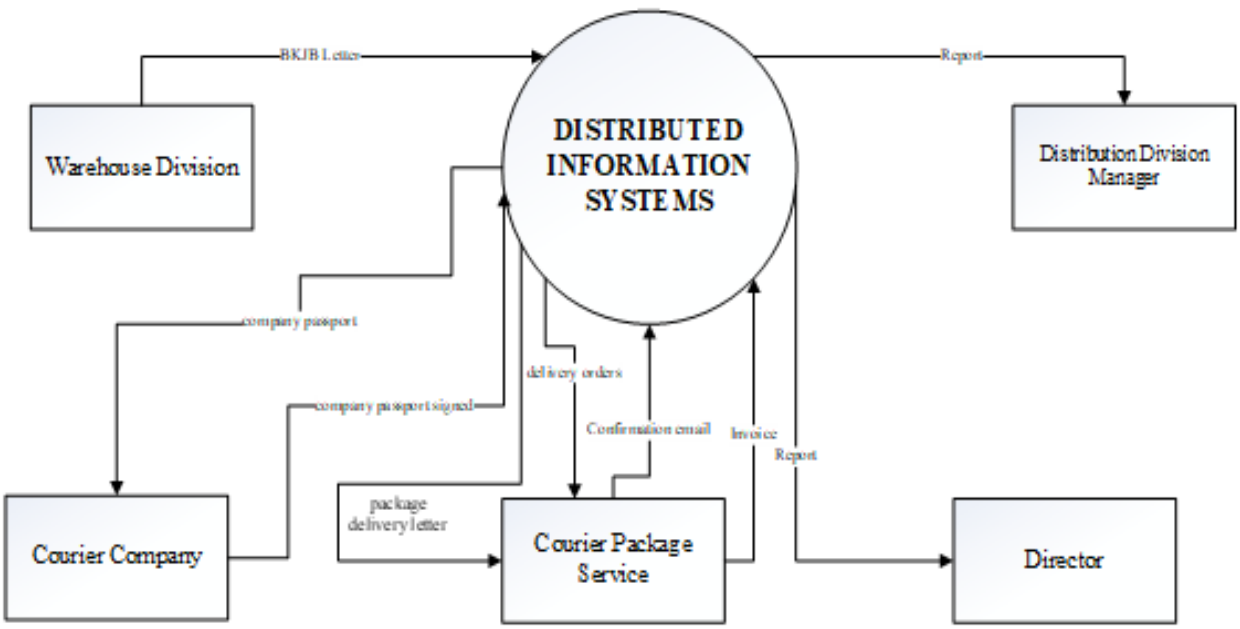


Fig. 1. Context Diagram of Web-Based Information System in Product Distribution

d. Implementation

This section briefly explains the use of the PT BCS distribution information system is the last stage in the results and discussion of the proposed system analysis. Some views of the software are described in the discussion this time, and the following software views. The main menu display of the distribution information system is a major display for doing every activity and processes. The main menu display can be seen in Figure 2.

Fig. 2. Main Menu Display

The company passport menu displays, adds, change and delete road mail data. The display of the company passport menu can be seen in Figure 3.

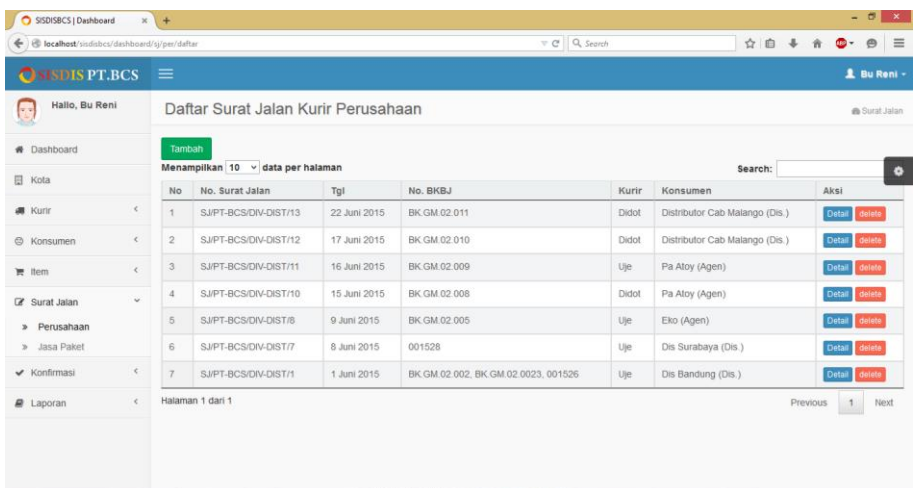


Fig. 3. Company Passport Display

4. Conclusion

Based on research that has been carried out through analysis and design, the proposed distribution information system can solve the problem of making a company passport, can store distribution data in the database, the process of confirming the shipment of products can be done by the recipient of the products themselves, and reporting can be done automatically so it doesn't take a long time.

Acknowledgements

Special thanks to PT. Bonli Cipta Sejahtera which has given time in cooperation for making this application and Universitas Komputer Indonesia that has supported this research activity.

References

Damayanti. (2020). *Sistem Informasi Pendistribusian Barang Bengkel Las dan Advertising Menggunakan Model SCM*. Jurnal Komputer dan Informatika. Vol 15 No 1. pp 209 -218

Ensour, H, S., and Alinizi, T, M. (2014). *The Impact of Manajemen Information System (MIS) Technologies on The Quality of Services Provided at The University of Tabuk.* Int. J. of Network Security & Its Applications. **6**(2)

Hartono, J. (2005). *Analisis dan Desain Sistem Informasi Pendekatan Terstruktur Teori dan Praktek Aplikasi Bisnis*. 3rd ed. Yogyakarta : ANDI.

Kadir, A., and Triwahyuni, T. (2013). *Sistem Informasi di Dalam Pengantar Teknologi Informasi*. Rev ed. Yogyakarta: ANDI.

Purnamasari, S, D., Maulana, and Fatoni. (2014). *Sistem Informasi Distribusi Barang Berbasis Web Service.* Seminar Nasional Teknologi Informasi dan Multimedia. STMIK AMIKOM Yogyakarta. pp 3.05-17 – 3.05-22

Sitanggang, A, S., and Kusumaningrum, S, V. (2019). *E-Tracking Application for Reporting Information System.* IOP Conf Series : Materials Science and Engineering.

Sulthoni, A., and Unang Achlison. (2015). *Sistem Informasi E-Commerce Pemasaran Hasil Pertanian Desa Kluwan Berbasis Web*. Jurnal Ilmiah Ekonomi dan Bisnis. **8** (1)

Theorin, A., Bengtsson, K., Provost, J., Lieder, M., Johnsson, C., Lundholm, T., and Lennartson, B. (2017). *An Event-Driven Manufacturing Information System Architecture for Industry 4.0.* International Journal of Production Research. **55**(5), pp. 1297-1311

Descriptive Statistics of Length of Stay in Hospital Wards: A Case Study at HCTM

Syazwan Md Yid [a,b], Rosmina Jaafar [a*], Seri Mastura Mustaza[a]

[a]Dept. Electrical, Electronic & Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[b]Medical Engineering Technology, Universiti Kuala Lumpur British Malaysian Institute, 53100 Gombak, Malaysia
*Email: rosmina@ukm.edu.my

Abstract

Information on hospital length of stay (LOS) of patients in the hospital ward is an important factor for planning and managing the resource utilization of a hospital. There has been considerable interest in controlling hospital cost and increasing service efficiency, particularly in surgical units where the resources are severely limited. The main objective of this paper is to analyze statistically patients' data for future LOS prediction models. A cross-sectional study was conducted on patients' data from 2015 to 2020 requiring intensive care unit (ICU) admissions in Hospital Canselor Tuanku Muhriz (HCTM), which is a teaching hospital in Malaysia. Factors that determined the LOS at the ICU were also explored by using multivariate regression analysis. It was found that general intensive care unit (GICU) patients were staying in the hospital at an average of 6 days across all age. However, patients of younger age (less than 12 years) requires longer ICU stay. Multivariate regression model shows only the category of the primary team handling the patients is the determinant for LOS. The diagnosis remarks in the database can be transformed into a code forms to allow better LOS prediction model.

Keywords: Length of stay; statistics; cross-sectional study

1. Introduction

Patient length of stay is one of important performance indicators of a hospital which gives a better understanding of the resource consumption and patients' flow through healthcare system (Panchami & Radhika, 2014). Researchers have found that there is a strong correlation between medical cost and length of stay (LOS) by studying the factors influencing medical cost (Luo, Lian, Feng, Huang, & Zhang, 2017). The productivity of hospitals drop significantly in two situations: First, if the hospital is in short supply for required resources such as facilities and manpower. Second, if the hospital is over equipped and the demand is less than the supply. Both of these situations occur due to significant fluctuations in hospital occupancy, which seriously restricts the efficient scheduling for resource allocation and management (Azari, Janeja, & Mohseni, 2012).

LOS in the ICU is one of the most important and influential factors in health financial management. Some studies considered LOS as a surrogate for hospital cost (Zhang & Liu, 2011). Analysis of determinants that can influence the length of ICU stay is of interest for both medical quality assurance and health economic aspects. Total LOS at the hospital and at the ICU varied with different care policies, different hospitals, and different countries (Gruenberg & D. A., 2006). A literature review mentioned that patients stayed on average 3.3 days at the ICU (Hunter, Johnson, & Coustasse, 2020). Nevertheless, the LOS could be different based on the condition of the patient and the location of the study.

In this article, we do a statistical analysis for possible determinants for predicting LOS and assess their suitability for planning resources, identifying unexpectedly long LOS, and benchmarking. This study used the general intensive care unit (GICU) census database recorded at Hospital Canselor Tuanku Muhriz (HCTM), Kuala Lumpur, Malaysia. This database is a system that records patients based on age, sex, race, diseases,

primary team, and LOS with similar resource utilizations, using information from the discharge summaries. This database is recorded by the head nurse in the GICU upon patient's admission.

The main aim of reviewing the ICU services in this article is to improve the efficiency of care through the determinants of LOS at ICUs and the factors associated with the LOS, using the existing database. A cross-sectional study of patients requiring ICU admissions was conducted as well.

2. Methodology

Data of patients who were warded at GICU and who were discharged from HCTM between January 1, 2015 and August 31, 2020, were included in the study. The characteristics of the patients admitted to the ICU were explored. The mean and standard deviation were calculated for continuous data, together with median. The proportion with 95% confidence interval (CI) was presented for categorical data. Skewed data were log transformed and checked for normality. Univariate analysis and multivariate regression analysis were conducted to identify the factors associated with the LOS of care at the ICU. All data were analysed and plotted using SPSS version 26.0. Analysis of mean difference between groups and univariate analysis are considered significant different if $p < 0.05$.

3. Results and Discussion

3.1. General statistics

Out of 4563 cases with ICU admissions, 4347 admissions with complete information were included. Table 1 shows the overall characteristic admissions data included in the analysis. Most of the admissions required care at the GICU are from Medical Team (42.3%), the rest are from Surgery Team (31.5%), Neuro Surgery Team (13.7%) and from other teams (12.5%) (Figure 1). The mean age of GICU admission was 54 years old; 50.7% of the admissions were of patients who were at least 57 years old or older. ICU utilization was higher among male patients: 60.2%.

Table 1. Characteristic of admissions involved in the study

	GICU stay (n = 4347)
Sex – (n[%])	
Male	2615 (60.2%)
Female	1732 (39.8%)
Age (years) – (Mean ± SD)	54.33 ± 18.4
Primary Team - (n[%])	
A&E	18 (0.4%)
Others	38 (0.9%)
Medical	1839 (42.3%)
Neuro Surgery	594 (13.7%)
O&G	127 (2.9%)
Oncology	17 (0.4%)
Orthopaedic	217 (5%)
SARI/PUI ICU	29 (0.7%)
Spine	75 (1.7%)
Surgery	1369 (31.5%)
Trauma	24 (0.6%)
Length of stay at GICU (days)	
Mean (SD)	6.54 (7.061)
Median (IQR)	4 (2-8)

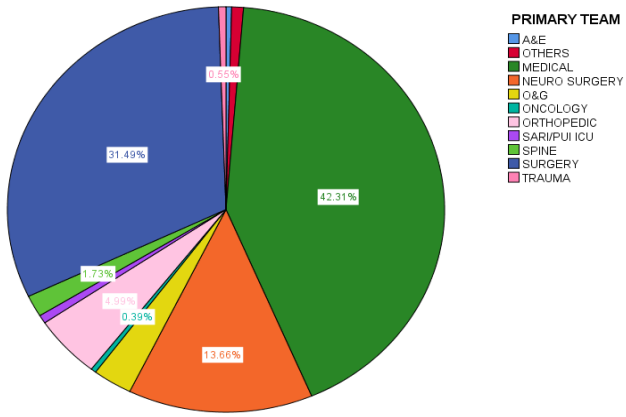


Fig. 1. Distribution of admission in GICU by type of primary team.

3.2. The Length of Stay at the GICU

Referring to Table 1, the total hospital stay for the 4347 admissions was 28,432 days. The mean of length of stay was 6.54 days (SD 7.061) with a median of 4 days; men required slightly longer ICU stays (0.33 day) but this difference was not statistically significant ($p = 0.129$). Of all the cases, the longest length of stay of 85 days was observed in the case of a 70 years woman. It was found that young children (below 12 years) required longer ICU days (mean 16.33 days for 12 years old patients). For patients above this age, the number of days at the ICU declined. The ICU stay gradually increased again above the age of 39 years. Results showed that there is a significant difference ($p < 0.001$) in average LOS at different primary team who handled each case with the highest at SARI/PUI ICU (10.07 days), followed by Oncology Team (9.71 days). ICU admission for obstetric reasons (O&G) had significantly shorter ICU length of stay (4.03 days) compared with the others.

Table 2. Univariate regression analysis for length of stay (LOS) at GICU (n = 4347)

		Length of stay at GICU (days)	
		Mean (SD)	**P - value**
Sex	Male	6.67 (6.884)	0.129
	Female	6.34 (7.317)	
Primary Team - (n[%])	A&E	7.83 (6.653)	< 0.01
	Others	5.55 (6.345)	
	Medical	7.14 (7.379)	
	Neuro Surgery	6.27 (5.81)	
	O&G	4.03 (6.119)	
	Oncology	9.71 (16.213)	
	Orthopaedic	6.51 (7.404)	
	SARI/PUI ICU	10.07 (8.124)	
	Spine	6.51 (6.456)	
	Surgery	5.97 (6.86)	
	Trauma	7.75 (8.451)	
Race	MELAYU	6.26 (6.710)	0.059
	CINA	6.95 (7.602)	
	INDIA	6.94 (7.224)	
	OTHERS	5.98 (6.843)	
	Age	-	0.201

Multiple regression analysis was performed by excluding the outliers in LOS and log linear transformation of LOS. Race, age and sex had no significant contributions in determining the length of ICU stay and were

excluded from the model. Table 2 shows that only the primary team handling the patients remained as major determinants for duration of stay at the ICU ($p < 0.01$). Given the same age, LOS varied by different primary teams.

From the regression analysis in this study, it was found that only the type of primary team handling the patients is the determining factors for LOS. We only managed to collect six main variables (age, gender, race, primary team, LOS and diagnosis). In this study, LOS at the ICU was 6.54 days (SD: 7.061 days), which was acceptable although this was a slightly longer length of ICU stay when compared with most other studies. The regression model lacks of many important variables. The GICU database lack of specific diagnosis as it is only stated as diagnosis remark. There were also missing information regarding the reasons of patients requiring the ICU, such as medical emergency, elective surgical, or emergency surgical cases. If the diagnosis remark is transferred to diagnosis code and reasons of patients requiring the ICU are available, it would be beneficial for better regression analysis.

4. Conclusion

This study has described the descriptive statistics of GICU patients' data who were warded in the HCTM. It was found that GICU patients were staying in the hospital at an average of 6 days across all age. However, patients of younger age (less than 12 years) require longer ICU stay. Multivariate regression model shows only the category of the primary team handling the patients is the determinant for LOS. For further work, the diagnosis remarks in the database can be transformed into a code forms to allow better LOS prediction model with high performance rate. The information discovered in this study is important as LOS is commonly used as the proxy indicator for hospital efficiency.

Acknowledgement

The authors would like to thank Ministry of Higher Education Malaysia for funding the study through the research grant TRGS/1/2019/UKM/01/4/3.

References

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting Hospital Length of Stay (PHLOS) : AAA multi-tiered data mining approach. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 17–24. https://doi.org/10.1109/ICDMW.2012.69

Gruenberg, D. A., Shelton, W., Rose, S. L., Rutter, A. E., Socaris, S., & McGee, G. (2006). Factors Influencing Length of Stay in the Intensive Care Unit. *American Journal of Critical Care*, 15(5), 502–509. doi:10.4037/ajcc2006.15.5.502

Hunter, A., Johnson, L., & Coustasse, A. (2020). Reduction of intensive care unit length of stay: The case of early mobilization. *Health Care Manager*, *39*(3), 109–116

Luo, L., Lian, S., Feng, C., Huang, D., & Zhang, W. (2017). Data mining-based detection of rapid growth in length of stay on COPD patients. *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, 254–258. https://doi.org/10.1109/ICBDA.2017.8078819

Panchami, V. U., & Radhika, N. (2014). A novel approach for predicting the length of hospital stay with DBSCAN and supervised classification algorithms. *5th International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2014*, 207–212. https://doi.org/10.1109/ICADIWT.2014.6814663

Zhang, A. H., & Liu, X. H. (2011). Clinical pathways: Effects on professional practice, patient outcomes, length of stay and hospital costs. *International Journal of Evidence-Based Healthcare*, *9*(2), 191–192. https://doi.org/10.1111/j.1744-1609.2011.00223.

A Web Based Early Warning Score Calculator and Data Logger for Patients Vital Signs Monitoring

Rosmina Jaafar[a*] and Nurul Izzati Kamdani[b]

[a]Dept. Electrical, Electronic & Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[b]Information Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
*Email: rosmina@ukm.edu.my

Abstract

An early warning score (EWS) of a person's vital signs is a guide often used by medical services to quickly determine the degree of health status of a patient. In a traditional EWS, vital signs are recorded manually on paper charts, making it cumber some to implement data retrievals. The advancement of computer technology offers fast computation, automation, and ease of data manipulation. It provides opportunity for ease of data retrieval. As such, we describe the development of a digital EWS calculator based on web applications. The system development utilized Laravel as a code editor, while phpMyAdmin was used for database management. Besides being a EWS calculator, the system also serves as a data logger for patient's vital signs monitoring. It requires a user to key in patients' vital signs into the system where all data are saved in a cloud data storage. The EWS system main functions are key-in data for vital signs (systolic and diastolic blood pressure, heart rate, body temperature, breathing rate and level of patient's conscious level referred as AVPU scale), display of data trend or history of vital signs, doctor's appointment, and surgery schedule for the patient that is registered into the system. The EWS system allows two types of users which are patient caretaker and super admin. All data and personal information are secured in the database. The EWS is important for patient health monitoring and future data retrieval will allow possible further analysis by the medical doctors when required.

Keywords: Early warning scores; vital signs; patients monitoring; data logger

1. Introduction

A systematic monitoring of the early warning score (EWS) from patients' vital signs has attracted the interest of many medical service providers worldwide beginning from the late 1990's after studies showed that in-hospital deterioration and cardiac arrest were often preceded by a period of increasing abnormalities in the vital signs (Morgan et. al, 1997). The standard vital signs from a human subject are systolic and diastolic blood pressure, heart rate, body temperature, breathing rate, oxygen saturation, and the level of patient' s consciousness commonly referred as AVPU scale ("alert, voice, pain, unresponsive"). A total score of five or more is statistically linked to increased likelihood of death or admission to an intensive care unit (Subbe et al, 2001). Since the awareness of the importance of tracking the EWS of a patient, few different EWS have been made available for special different situations. The common EWS are briefly described as follows (Doyle, 2018):

- National Early Warning Score (NEWS) for general patient monitoring based on the Royal College of Physicians set point
- Modified Early Warning System (MEWS) for general patient monitoring without the need to input oxygenation information
- Pediatric Early Warning Score (PEWS) for pediatric population that also focuses on child behaviors

- Early Warning Score for Obstetrics (EWSO) for obstetricians and neonatologists to monitor health status of maternal-fetal during antenatal phase

In the early days, all EWS systems evolved from the recording of the related patients' vital signs on paper chart which is cumbersome. Many efforts have been made to transform EWS into electronic systems or computerized ones for ease of data handling and retrieval. Over the years, there have been quite a number of electronic calculators of EWS available in the market. One of such systems is MDCalc, which is a free online medical reference for healthcare professionals that provides point-of-care clinical decision-support tools, including medical calculators, scoring systems, and algorithms. This MDCalc is also available in mobile and web applications. Nevertheless, the back-end data for MDCalc is not easily accessible for users who want to have a full access to the recorded data.

As such there is a need for developing a custom-made EWS system such that the system not only can serve as the EWS calculator, but also as a data logger system especially for researcher. A custom-made system can be designed tailored to the specific objectives as stipulated by the researcher. This paper describes the development of a EWS system that follows the specification of MEWS with additional input data of diastolic blood pressure. The EWS system also provides a learning platform for the programmer to implement web applications system development.

2. Methodology

2.1. EWS System Design

A flow chart is a diagram that explain the flow of activities that is carried in a system. Fig. 1 shows the flow chart for the Caretaker in handling the embedded EWS calculator and data logger for patient's vital signs monitoring.

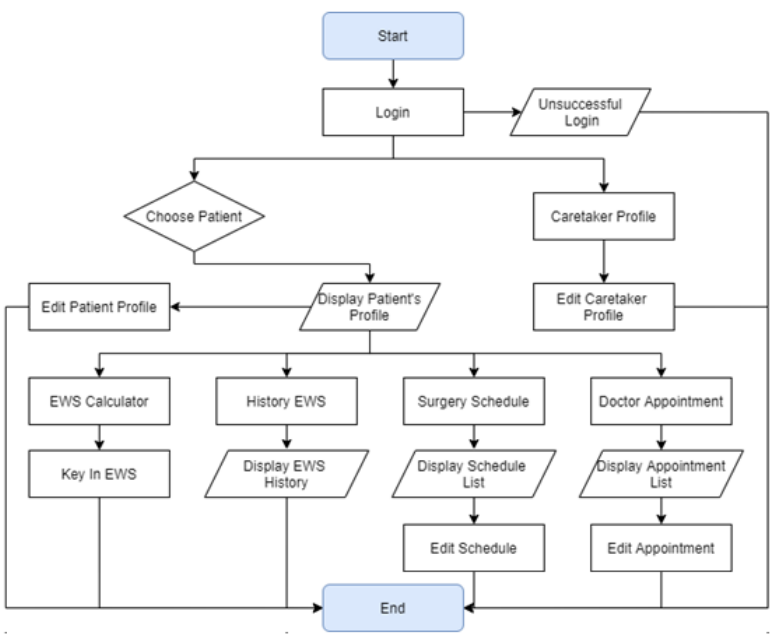


Fig. 1 Flow chart for handling the EWS system

2.2. Entity Relationship Diagram

The entity relationship diagram (ERD) is a way to describe and view the relationship between the entities involved. The use of ERD greatly helps the system developer to understand and see an overview of the whole system to be developed. Fig. 2 shows the entity relationship model of the EWS System.

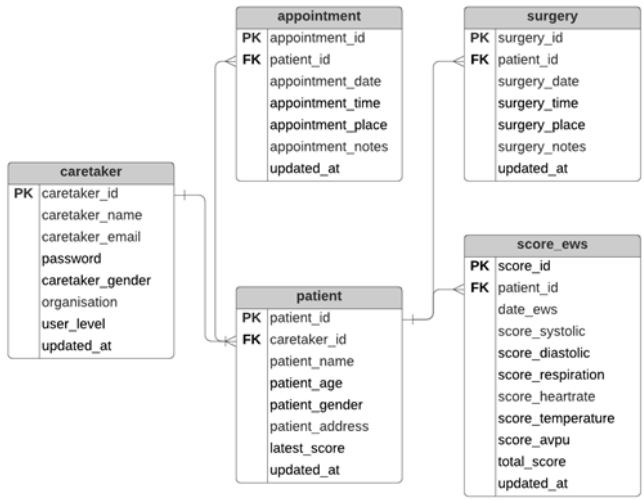


Fig. 2 Entity-Relationship Diagram for Patient Monitoring System

2.3. Data Dictionary

A data dictionary is a set of information that describes the content, format, and structure of a database and the relationships between its elements. It is used to control data access and database manipulation. In the data dictionary, each data attribute will be assigned a data type, size, key and a brief description of the attributes used in this system database.

Laravel, which is a web application framework with expressive and elegant syntax, was chosen for developing the web applications for the EWS system. It was chosen because it is a free, open source (Anonymous, 2017) modular packaging system with a dedicated dependency manager. The advantage of using Laravel include it being easy in accessing relational databases, having utilities that aid in application deployment and maintenance, and its orientation toward syntactic sugar (Martin, 2015).

2.4. System Interface

The interface design of a system plays an important role in determining the usability of the system. A good user interface should be easy and quick for users to learn. In addition, the interface also serves as a medium of interaction between users and the system. The system interface includes user login, patient registration, vital signs data input, patient history, doctor's appointment and surgery schedule. The system has two types of users: Caretaker and SuperAdmin. All user interfaces for Caretaker are accessible by the Super Admin who has the sole function of downloading patient data and save the data in excel spreadsheet. Super Admin can also view all Caretaker's activity.

The phpMyAdmin which is another free software tool was utilized to handle the administration of MySQL over the web. This software supports a wide range of operations on MySQL to be performed via the user interface. This includes managing databases, tables, columns, relations, indexes, users, permissions, etc.

3. Results and Discussion

3.1. EWS System Prototype Testing and Evaluation

Development of the first prototype of the web based EWS calculator and data logger for EWS was completed. The program developer has tested system functionality few times using black box where the overall score for functionality tests was 100% as it passed all tests conducted. After the EWS system completed its development process, it was uploaded to the server. The EWS prototype is functioning well based on system functionality test. Full user satisfaction testing is yet to be implemented. A pilot study of patients' vital signs monitoring via EWS recording is being planned to take place soon involving data recording on real patient monitoring. The EWS system prototype can be modified and improvised based on user recommendation from the pilot study, to make it meet all user requirement and specifications.

3.2. Recommendation for Future System Enhancement

The EWS system has many rooms for system enhancement. One of it is to add an alert function such that when a patient shows vital signs deterioration, the EWS triggers a warning so that necessary patient care can be provided as soon as possible.

4. Conclusion

In conclusion, a custom-made digital system of EWS was successfully developed. The EWS system allows patient vital signs to be recorded and calculated the early warning scores which contain valuable patient health information. The EWS data are saved for future data retrieval and the trend or history of EWS also provide useful information for clinical decision support. Besides the vital signs and EWS data, the EWS system also includes patient's doctor appointment, and surgery schedules for ease of patient's clinical care management.

Acknowledgements

The authors would like to thank Ministry of Higher Education Malaysia for funding the work through the research grant TRGS/1/2019/UKM/01/4/3.

References

Anonymous. (2017). The real-time community site Voten goes open source. Laravel News. June 16, 2017. Retrieved April 30, 2021.

D. John Doyle. (2018). Clinical Early Warning Scores: New Clinical Tools in Evolution. The Open Anesthesia Journal, 12, 26-33.

Martin Bean. (2015). Laravel 5 Essentials. Birmingham, UK: Packt Publishing Limited.

Morgan, R., Lloyd-Williams, F., Wright, M., Morgan-Warren, R.J. (1997). An early warning scoring system for detecting developing critical illness. https://www.scienceopen.com/document?vid=28251d22-8476-40a6-916d-1a34796816e4

Subbe, C.P., Kruger, M. and Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. Quarterly Journal of Medicine. 94 (10): 521–6. doi:10.1093/qjmed/94.10.521. PMID 11588210.

Anaesthetist Rostering Web Application for Hospital Canselor Tuanku Muhriz

Norizal Abdullah[a]*, Masri Ayob[a], Nurul Camelia Murad[b]

[a]Data Mining and Optimization Lab, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
[b]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
* Email: norizal.abdullah23@gmail.com

Abstract

Hospital Canselor Tuanku Muhriz (HCTM) is one of the hospitals that provide medical, surgical treatment and nursing care for patients. An anaesthetist is a specialist doctor that is responsible for providing anaesthesia to patients which are mainly used in the surgical department. The hospital's administration must ensure that an anaesthetist is available to treat the patients. Making a roster for anaesthetists is difficult because it is tied to their location, work shift, and workstation when using a manual method. This project aims to create a web application that can be used to control and manage anaesthetist scheduling in real-time. PHP is the programming language, MySQL is the database, and Laravel is the web application platform for this method.

Keywords: Anaesthetist Rostering; Hospital; Web Application

1. Introduction

Currently, the anaesthetist rostering in Hospital Canselor Tuanku Muhriz (HCTM) uses a manual system using Google Form, Microsoft Excel and Microsoft Word to collect the requests and organize the roster. This manual method appears to be ineffective, difficult to maintain, and inefficient. Therefore, this work proposes the design and development of a web application for an anaesthetist rostering at HCTM. The purpose of this system is to help the roster maker to arrange the roster for anaesthetists.

The web application must be able to manage the roster planning regarding the workstation demand for the anaesthetist and the request for every anaesthetist. In the next section, we begin with the literature review and the methodology for an anaesthetist rostering web application. Next, we present the result of the development and end it with the conclusion.

2. Literature Review

Pressure, nausea, vomiting, and ill may occur after surgery, as well as stress-induced catabolism, decreased pulmonary function and increased cardiac demands..These issues can lead to complications, hospitalization, postoperative exhaustion, and a longer recovery time. By providing minimally invasive anaesthesia and pain relief, as well as working with surgeons, surgical nurses, and physiotherapists to minimize risk and pain, the anaesthetist plays an important role in promoting early postoperative recovery (Kehlet & Dahl, 2003). In the last decade, the reach of anaesthetists' work in hospital practice has broadened. Anaesthetists are experts in emergency medicine, intensive care, and the treatment of acute and chronic pain. Some anaesthetists may have testing, teaching, or administrative responsibilities (Kinzl, Knotzer, Traweger, Lederer, Heidegger, & Benzer, 2005).

The assigning of tasks to employees through a schedule is known as duty scheduling. It is the method of analysing an organization's workload, the time available to execute the workload, and the task allocation based on the available time (Yange, Onyekwere, Okeke, & Applications, 2020). Staff rostering, or the preparation of

work schedules in healthcare organisations is a complex and time-consuming activity that affects healthcare workers all over the world daily. It's especially difficult because different personnel requirements exist on different days and shifts, resulting in a variety of constraints (Zhu, Tong, Low, Lau, Chen, & Wang, 2019).

Personnel scheduling, also known as rostering, is the method of creating work schedules for employees so that a company can meet the demand for its products or services (Ernst, Jiang, Krishnamoorthy, & Sier, 2004). The majority of healthcare workers do use a manual system to keep track of their schedules. The schedulers must also be mindful of the doctors who are on call and who are on vacation. An operation may be conducted without a scheduled anaesthetist due to a scheduling error (Scholiadis, du Toit, & Sevel, 2005). Overestimation of operational time results in unused operating rooms, whereas underestimation results in unplanned extra work or case cancellation, all of which may raise costs (Wright, Kooperberg, Bonar, & Bashein, 1996).

3. Methodology

3.1. The Existing System

HCTM uses a manual technique to generate an anaesthetist rostering. The roster is prepared for every month and week. Mostly, the whole process is done by the roster maker which is the head of the anesthesiology department.

3.2. The Proposed System

The proposed system is the anaesthetist rostering web application. This web application will prepare the roster for the anaesthetist into shifts based on workstation demand and requirement. This system generates a new duty roster every month. The roster maker can manage the workstation demand and next the system can generate the roster for the anaesthetist.

3.3. Use Case Diagram

The use case diagram for the proposed system as shown in Fig. 1. depicts the actors (anaesthetist and admin) and their interactions with the system.

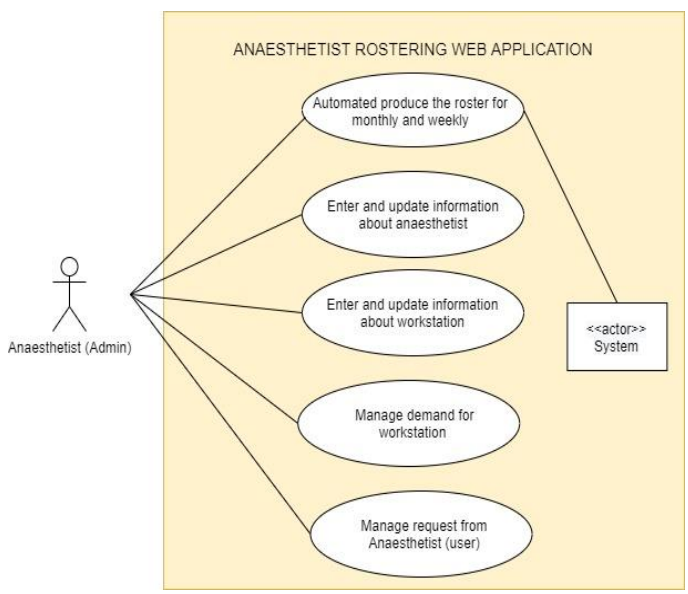


Fig. 1. Use Case for anaesthetist rostering web application

4. Results

The system was evaluated to ascertain its compliance with the requirements. The sample outputs of the newly proposed system are shown in Fig.2 and 3. Fig. 2. show the roster page for the anaesthetist rostering web application. This page extracts the anaesthetist schedule based on their level and workstation. Fig. 3. show the interface for the workstation demand. This page is used to manage the workstation demand for the roster.

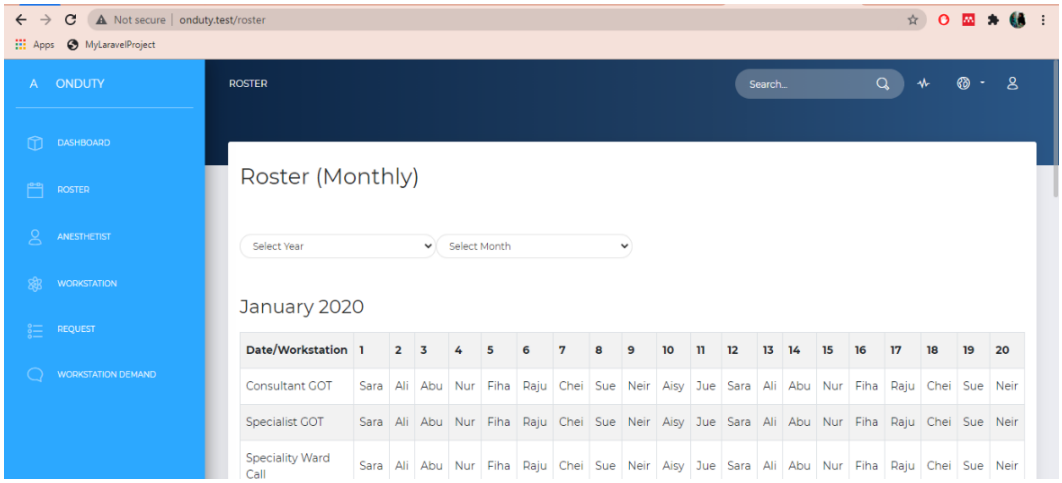


Fig. 2. Interface for roster

5. Conclusion

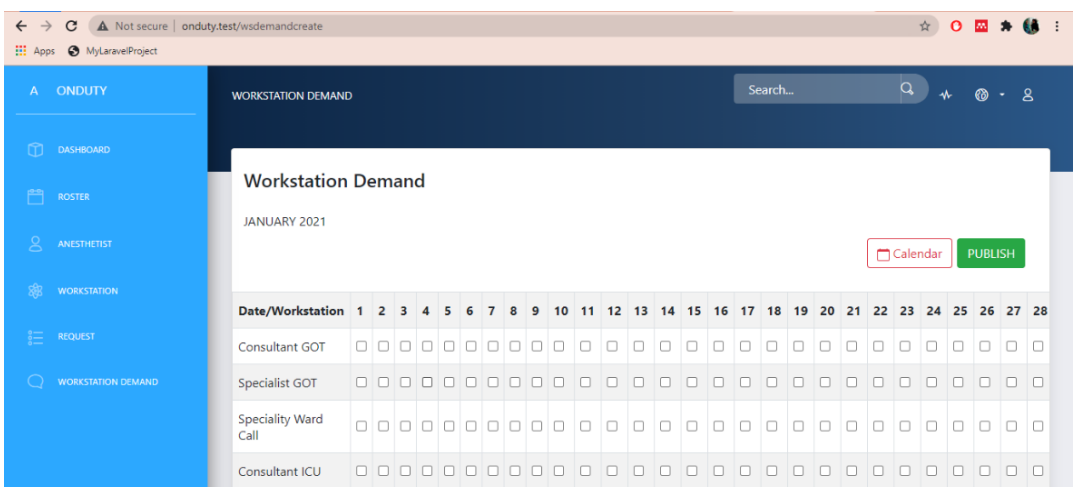


Fig. 3. Interface for workstation demand

In a nutshell, this system is still in the development phase and soon will be integrated with the artificial intelligence engine (to automatically construct the anaesthetist roster using a meta-heuristic approach) and tested on a real-life anaesthetist roster at HCTM. Research carried out shows that the computerized system yields more advantages than the manual system of rostering (Paschou, Papadimitiriou, Nodarakis, Korezelidis, Sakkopoulos, & Tsakalidis, 2015). Further works are still required to make the system have more functionality.

Acknowledgements

The authors wish to thank the Universiti Kebangsaan Malaysia and the Ministry of Higher Education Malaysia for supporting and funding this work (grant ID: TRGS/1/2019/UKM/01/4/1).

References

Ernst, A. T., Jiang, H., Krishnamoorthy, M., & Sier, D. J. E. j. o. o. r. (2004). Staff scheduling and rostering: A review of applications, methods and models. *153*(1), 3-27.

Kehlet, H., & Dahl, J. B. J. T. L. (2003). Anaesthesia, surgery, and challenges in postoperative recovery. *362*(9399), 1921-1928.

Kinzl, J. F., Knotzer, H., Traweger, C., Lederer, W., Heidegger, T., & Benzer, A. J. B. J. o. A. (2005). Influence of working conditions on job satisfaction in anaesthetists. *94*(2), 211-215.

Paschou, M., Papadimitiriou, C., Nodarakis, N., Korezelidis, K., Sakkopoulos, E., & Tsakalidis, A. (2015). Enhanced healthcare personnel rostering solution using mobile technologies. *Journal of Systems and Software, 100*, 44-53. doi:https://doi.org/10.1016/j.jss.2014.10.015

Scholiadis, T., du Toit, P. W., & Sevel, D. (2005). *Web Drugs: Anaesthetics Automated Scheduling System*. Retrieved from

Wright, I. H., Kooperberg, C., Bonar, B. A., & Bashein, G. J. T. J. o. t. A. S. o. A. (1996). Statistical modeling to predict elective surgery time: comparison with a computer scheduling system and surgeon-provided estimates. *85*(6), 1235-1245.

Yange, T. S., Onyekwere, O., Okeke, O. B. J. J. o. C. S., & Applications. (2020). A Duty Scheduler for Veterinary Teaching Hospitals Using Tabu Search Algorithm. *8*(1), 30-39.

Zhu, L. R., Tong, J. H., Low, M. Y. H., Lau, B. H., Chen, M. X., & Wang, Z. (2019). *An Automated Staff Roster Planning System (SRPS) For Healthcare Industry.* Paper presented at the 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

Anaesthetist Rostering Mobile Application for Hospital Canselor Tuanku Muhriz

Norizal Abdullah[a]*, Masri Ayob[a], Raja Nur Natasha Raja Ahmad Anuar[b]

[a]Data Mining and Optimization Lab, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
[b]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
*Corresponding Author Email: norizal.abdullah23@gmail.com

Abstract

Hospital Cancelor Tuanku Muhriz (HCTM) is a place that provides health care services to society. A lot of health-related cases have been referred to HCTM especially at the Surgical Department. In the surgical department, an anaesthetist is a specialist doctor responsible for providing anaesthesia to the patients before or after the surgery phases. The hospital manager must ensure that the anaesthetists must always be available during the surgery phases. Rostering anaesthetists is challenging because it needs to fit with the anaesthetist's preferences and requests, type of duty, the demand of the workstation and rostering planning. The process to balance the workload among anaesthetists for job satisfaction is hard with the manual system. Thus, this work designs and develop a mobile application that can be used to manage the anaesthetist rostering in HCTM. The mobile application has the functions to allow anaesthetist to make a roster request, view roster, and can request a swap roster, accept or reject swap roster request. These functions can give more flexibility to anaesthetist work at their preferred time and day.

Keywords: Anaesthetist Rostering; Hospital; Mobile Application

1. Introduction

All hospital should deliver excellent and fast services to patients. Due to that, hospital managers must ensure their rostering process can run smoothly and efficiently to avoid the absence of staff on duty and excessive increase in the hospital workload which can cause physical fatigue in healthcare (Güler & Geçici, 2020). In Hospital Canselor Tuanku Muhriz (HCTM), an anaesthetist is an important person who must be available during the surgery phase in the Surgical Department. Currently, the anaesthetist rostering in HCTM uses manual systems such as Google Form and Microsoft Excel to collect the requests, set up the roster planning and organize the roster. This manual system seems unreliable, hard to manage and lead the manual systems inefficient.

Therefore, this work aims to design and develop a mobile application for anaesthetist rostering. The mobile application must be able to allow an anaesthetist to make a roster request, view roster, and can request a swap roster, accept or reject swap roster request etc. In the next section, we begin with the literature review and continue with the methodology for design and develop the anaesthetist rostering mobile application. Next, we show the result of our development and end it with the conclusion.

2. Literature Review

An anaesthetist is a physician that has many tasks. They are responsible for administering anaesthetic or sedation during medical procedures (Aghsaei, 2014). Preparing a roster for anaesthetists in a hospital is a complicated and time-consuming task. A Schedule often called a roster, is a list of employees and associated information (i.e., location, working times, responsibilities for a given period for a week, month or season)

(Yange et. al., 2020). The issue of rostering anaesthetist is challenging due to the multiple demands and requests from the hospital and anaesthetists. Duty scheduling is the assignment of tasks to staff. It is the process of analysing the workload in an organization, the time available to implement the workload and the distribution of the work according to the time available (Yange et. al., 2020). This task is beyond the capability when the roster maker needs to fulfil anaesthetist requests and always need to keep updating the roster.

Paschou et. al., (2015) proposed an intelligent mobile device application, to deal with lots of obstacles in day-to-day medical rostering management and improved workflow in medical units. A few functions were implemented to allow health care to get informed online and make ad-hoc changes to their shifts dynamically through smartphone (Paschou et. al., 2015). Besides that, they implemented a calendar function to increase the efficiency of the application in managing the roster. In other work, Zhu et. al., (2019) enhanced the existing approach of manual staff roster planning by significantly reducing the number of man-hours used in the process of planning and minimizing the possibility of human error in the rostering process.

3. Methodology

3.1. The Existing System

HCTM uses a manual system to generate a roster to schedule an anaesthetist. The monthly roster request is prepared every month to collect the request for duty using Google Form. Anaesthetists can choose their duty on preferred days and shift in that month. The roster maker will manage the roster by taking the data from the Google Form and organize it inside Microsoft Excel.

The things that always requested by anaesthetist are:

- Leave
- Workshop/Conference
- Meeting
- Dissertation
- No Call
- AM Shift
- PM Shift
- Others

3.2. The Proposed System

The proposed system is a mobile application. This mobile application will help to collect the request from the anaesthetist. The type of requests that are implemented in this application is a Roster Request, Emergency Request and Swap Roster Request. Roster Request is used to collect monthly requests from anaesthetists before the schedule is released. Therefore, the anaesthetist can plan their schedule wisely for that month. Sometimes, anaesthetists will experience an unavoidable situation or problem that causes them unable to work that day or time. Emergency Request can be a good platform for anaesthetists to request or inform the hospital about their situations. This application allows anaesthetists to exchange duty with their colleagues through Swap Roster Request. This function can give more flexibility to anaesthetist work at their preferred time and day.

3.3. Use Case Diagram

The use case diagram for the proposed system as shown in Fig. 1. depicts the interaction between the system and the actors (anaesthetist and admin).

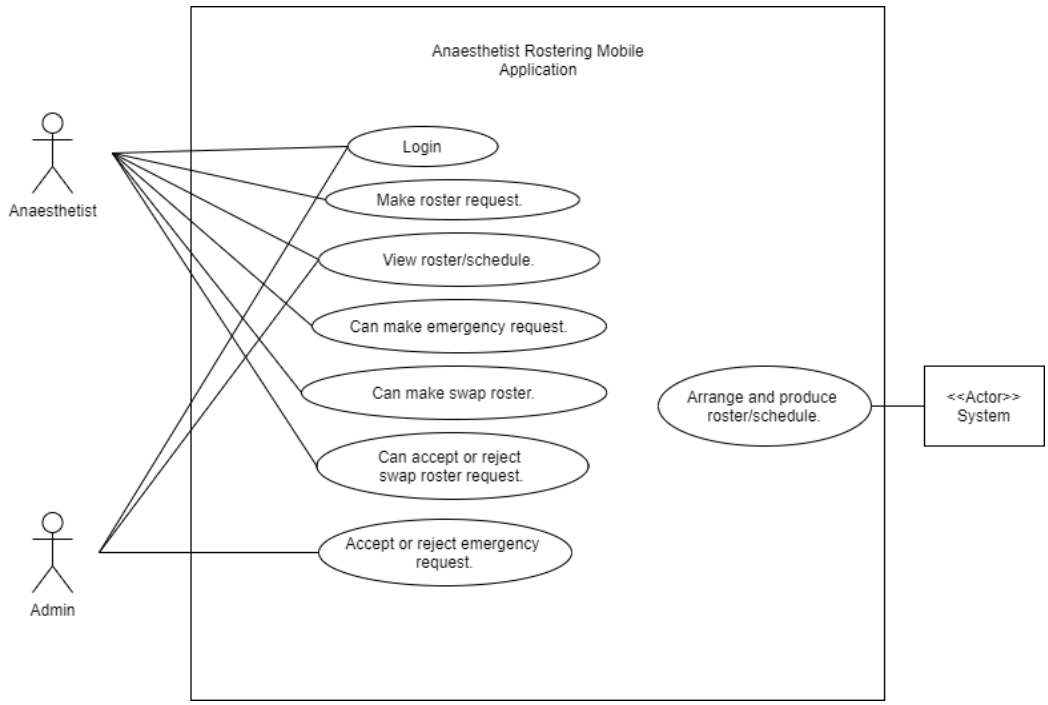
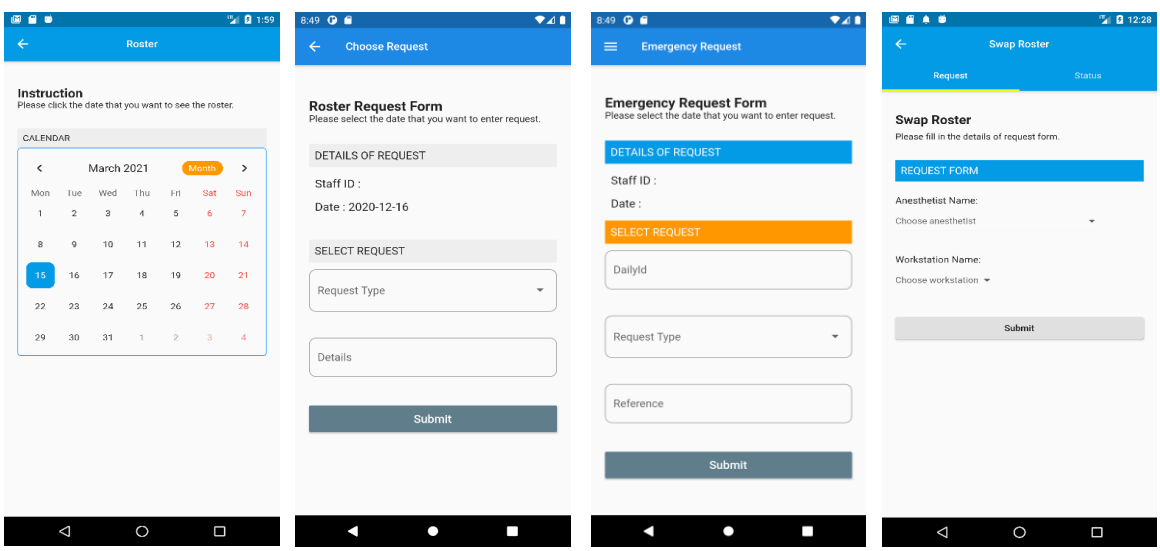


Fig. 1. Use Case for System Design

4. Results

This section will show the user interface for anaesthetist rostering mobile application. This application was developed using PHP and Dart as the programming language and MySQL as the database. Fig. 2. show the



(a) (b) (c) (d)

Fig. 2. (a) Interface for roster; (b) Interface for request form; (c) Interface for emergency request form; (d) Interface to make swap request

main interface for anaesthetist rostering mobile application which is consists of the roster, request form, emergency request form and swap request.

5. Conclusion

An anaesthetist rostering mobile application that had been presented in this work is not been testing yet because it still under design and development. The aims is to assist the anaesthetist rostering in HCTM by enabling real-time communication between anaesthetists through smartphone and tablets. The features that are being implemented in this system is to give flexibility to the admin, that is the roster maker, to easily manage the request for anaesthetist duty and access the roster from any location. Also, this application can help reduce the workload for the roster maker that is responsible to manage the roster. Further enhancements are still required to make the application work better. Additionally, the proposed mobile application also can be customized to meet the needs of different healthcare facilities that may have different demand from such an integrated system. The mobile application will be integrated with the web-based application and the artificial intelligence engine to automatically construct the anaesthetist roster based on the request made by the anaesthetist using the Anaesthetist Rostering Mobile Application.

Acknowledgements

The authors wish to thank the Universiti Kebangsaan Malaysia and the Ministry of Higher Education Malaysia for supporting and funding this work (grant ID: TRGS/1/2019/UKM/01/4/1).

References

Aghsaei, S. (2014). *Anesthesiologist and nurse anesthetist (CRNA) assignment on the day of surgery*: Northeastern University.

Güler, M. G., & Geçici, E. (2020). A decision support system for scheduling the shifts of physicians during COVID-19 pandemic. *Computers & Industrial Engineering, 150*, 106874. doi:https://doi.org/10.1016/j.cie.2020.106874

Paschou, M., Papadimitiriou, C., Nodarakis, N., Korezelidis, K., Sakkopoulos, E., & Tsakalidis, A. (2015). Enhanced healthcare personnel rostering solution using mobile technologies. *Journal of Systems and Software, 100*, 44-53. doi:https://doi.org/10.1016/j.jss.2014.10.015

Yange, T. S., Onyekwere, O., Okeke, O. B. J. J. o. C. S., & Applications. (2020). A Duty Scheduler for Veterinary Teaching Hospitals Using Tabu Search Algorithm. *8*(1), 30-39.

Zhu, L. R., Tong, J. H., Low, M. Y. H., Lau, B. H., Chen, M. X., & Wang, Z. (2019). *An Automated Staff Roster Planning System (SRPS) For Healthcare Industry.* Paper presented at the 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

Investigating Feature Relevance for Essay Scoring

Jih Soong Tan[a]*, Ian K. T. Tan[b]

[a]Priority Dynamics Sdn Bhd, One City, Subang Jaya and 47650, Malaysia
[b]Monash University Malaysia, Bandar Sunway, Subang Jaya and 47500, Malaysia
*Email: jsoong@prioritydynamics.com

Abstract

Human grading of essays requires significant effort that is time consuming and vulnerable to be biased to the varying human graders. There has been numerous research effort in recent years on automated essay scoring (AES). The majority of the researches are based on extracting multiple linguistic features and using them to build a classification model for essay scoring. There are 3 main groups of features that are commonly being investigated for AES, namely lexical, grammatical, and semantic features. In this paper, we conducted empirical studies to investigate the influence of the different groups of features on the accuracy of the AES classification models based on a commonly used approach for AES research. The results exposed that the semantic feature, prompt, is the weakest group among the feature groups and this is due to the typical overfitting of the classification model when using the essay prompt.

Keywords: Auto Essay Scoring; Features; Importanc; EASE; ASAP

1. Introduction

Essays are generally used in academic writing which determines the understanding of students based on their arguments. However, in order to grade these essays, the effort needed by human graders will require time to ensure fair assessments. This is because human grading is vulnerable to be biased and will vary depending on the events that precede the human grader's life (Shermis& Burstein, 2003). An automated essay scoring (AES) computing system is ought to be capable of overcoming all these human graders' shortcomings by being consistent and fair throughout the essay evaluation (Shermis& Burstein, 2003; Janda et al., 2019). As far back as 1966, Page (1966) first invented an AES system called Project Essay Grade (PEG). Since then, there have been innovations and new systems in the AES field such as a newer version of PEG (Page, 1994), e-rater V2 (Attali& Burstein, 2006), and IntelliMetric (Elliot, 2001). Among all these systems, the linguistic features can be grouped into 3 groups of features, which are lexical, grammatical and semantic features.

In the previous study by Shermis& Burstein (2003), they have reported that the key properties of a good essay are written around the given prompt, well-structured, smooth flow, good grammar application, length, good spellings, and punctuation. Hence, we propose feature influence study to find the weak points of current feature engineering using a generic approach of feature engineering for AES for potential further improvement in addressing the AES classification accuracy. Using known state of the art learning algorithms for the classification models, the most influential and the least influential or the weak point of the current feature engineering method is discovered.

2. Related Work

For feature engineering in AES, there has been several efforts done by the other researchers. Phandi et al.(2015) have worked on AES by implementing the Enhanced AI Scoring Engine (EASE)* engine to extract

*https://github.com/edx/ease

features and using the feature to train a Bayesian Linear Ridge Regression (BLRR) model which scored an average of 0.7045 Quadratic Weighted Kappa (QWK) score. The EASE engine groups feature into 4 groups which are length, part-of-speech (PoS), bag of words (BoW), and prompt. It is often being used by others as the baseline feature engineering comparison to their own research projects (Eid & Wanas, 2017; Latifi&Gierl,2020; Janda et al., 2019) as it is invented by one of the top 3 winners of the Automated Student Assessment Prize (ASAP) competition. Hence the EASE engine is considered to be a robust and baseline engine for AES.

Coh-Metrix is a system proposed by Graesser et al. (2004) where it is an integration of varying software modules that extract features based on language, discourse, cohesion, and world knowledge (McNamara et al., 2010). Latifi & Gierl (2020) have taken the ASAP dataset and built a random forest model based on Coh-Metrixfeatures, which scored an average of 0.7 QWK score.

Eid &Wanas (2017) have proposed to focus on lexical features for AES, where they gathered 22 lexical features from three other pieces of research on lexical features, which scored an average of 0.684 QWK score. Janda et al. (2019) proposed 3 main groups of features; syntactic, semantic, and sentiment, that consists of 30 features and they worked on several feature selection techniques to select the top features. These were then use for a classification model of a three-layer neural network that resulted in an average QWK score of 0.793.

From our review of other related work, the results reported by Phandi et al. (2015) using the EASE feature extractor is a good baseline for investigation. We propose to base the evaluation on Phandi et al. in order to identify the influential feature groups.

3. Evaluation Methodology

3.1. Data preprocessing

We use the set 2 from the dataset released for the ASAP competition so that we can focus on our investigation. We extract the features from the dataset by using the functions provided by EASE. Features generated by EASE group into four feature groups including length, part of speech (PoS), prompt, and Bag of Words (BoW) refer to Phandi et al's (2015) paper. We followed Phandi et al. (2015) method of data preprocessing to get as closest as the results they get with EASE.

3.2. Learning algorithm and evaluation metric

Multinomial Naïve Bayes (NB) was added on top of the learning algorithms Support Vector Machine (SVM) and BLRR that were applied in Phandi et al. (2015). NBis known to be suitable for multinomial distributed data for short text classification. As reported byPhandi et al. (2015), BLRR has been often proven to provide good results in natural language processing tasks. SVM regression is selected as the comparison against BLRR.The implementation of the learning algorithms is written in the Python programming language (version 3.8) utilizing the Python `scikit-learn` library.

To evaluate the trained models, QWK was used to calculate the agreement between two raters, the human rater and the trained models. It considers the possibility of the agreement happening by chance (Vanbelle& Albert, 2009). QWK is the official evaluation metric being used for the ASAP competition. Also, the work by Phandi et al. (2015), Latifi&Gierl (2020); Janda et al. (2019) use QWK for their evaluation.

3.3. Experimental Setup

The pre-processed data will be duplicated into 4sets for the purpose of generating "Exclude one feature group" dataset where each of these new datasets will exclude a feature group each, and these are referred to as

"Exclude Length", "ExcludePoS", "ExcludeBoW"and "Exclude Prompt". A set of pre-processed data will be kept for comparison, referred to as "All features". The 5 training datasets will be used to train the three models separately. Then, we predict the scores based on the test sets. The predicted scores will be taken into the QWK evaluation metric to compute the agreement between the human rater's scores and the AES predicted scores.

4. Results and Discussion

4.1. QWK scores result for comparison

The QWK scores for the "All features" dataset and 4 "Exclude one feature group" datasets were computed for the three trained models. The trained models' results are summarized in Table 1. "All feature group" shows the BLRR model outperforming the rest of the models, which is in agreement with Phandi et al. (2015). For "Exclude one feature group", the most influential sets are bold-faced, and the least influential sets are underlined. The length feature is the most influential feature group. However, the prompt feature seems to be lacking. The QWK score of "Exclude prompt" in SVM and BLRR compared to "All features" show it is overfitting the trained model. By overfitting, it means the prompt feature has worsened the models.

Table 1. Results for all EASE features and except one feature group.

Feature Group	Features Used	QWK Score		
		NB	SVM	BLRR
All feature group	All Features	0.517	0.601	**0.626**
Exclude one feature group	Exclude Length	**0.444**	**0.565**	**0.601**
	Exclude PoS	0.511	0.583	0.617
	Exclude BoW	<u>0.546</u>	0.599	0.604
	Exclude Prompt	0.494	<u>0.636</u>	<u>0.657</u>

The EASE function uses the Natural Language ToolKit (NLTK) to tokenize the essay topic into prompt words. Subsequently, it finds the synonym of prompt words through the WordNet corpus in NLTK. Then, it counts the synonym of prompt words and prompt words. We postulate that the reason for the prompt features to be the least influential and to over fit is due to its weakness of extracting the semantic attributes. Semantic attributes correspond to the contextual meaning of words or a set of words (Janda et al., 2019). It is crucial for essay evaluation that the essay is written around a prompt or essay topic semantically (Norton, 1990). Hence, we believe the EASE engine took into consideration of all PoS, which caused the prompt feature to overfit. PoS such as conjunctions and adpositions do not bring any contextual meaning, which could add noise to the dataset.

Also, the method EASE applied to extract the semantic attributes is too brief and can be further improved in the future. It only takes into consideration the separate words instead of a pair of words or sentences, which makes it unable to capture context where a sentence or essay is starting to digress. As reported by Miltsakaki&Kukich (2000), coherence between a pair of words or sentences is the key to make text semantically meaningful.

5. Conclusion

We have experiments to investigate the weak point of the generic approach of feature engineering in AES. We propose to compare the four types of features extracted from EASE by using the "Exclude one feature group" datasets, then compare their QWK score with the "All features" set. As the comparison between the sets, our work has shown that the prompt feature is the weakest feature among the four types of features. The

"Exclude one feature group" set has represented that the QWK scores of SVM and BLRR without prompt features are better in performance. Hence, we showed that the prompt feature is overfitting in the dataset. As such, we can work on researching the new prompt feature in the future.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3).

Eid, S. M., &Wanas, N. M. (2017, November). Automated essay scoring linguistic feature: Comparative study. In 2017 Intl Conf on Advanced Control Circuits Systems (ACCS) Systems & 2017 Intl Conf on New Paradigms in Electronics & Information Technology (PEIT) (pp. 212-217). IEEE.

Elliot, S. (2003). IntelliMetric: From here to validity. Automated essay scoring: A cross-disciplinary perspective, 71-86.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers, 36(2), 193-202.

Janda, H. K., Pawar, A., Du, S., & Mago, V. (2019). Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation. IEEE Access, 7, 108486-108503.

Latifi, S., &Gierl, M. (2020). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. Language Testing, 0265532220929918.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., &Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4), 292-330.

Miltsakaki, E., &Kukich, K. (2000). Automated evaluation of coherence in student essays. In Proceedings of LREC 2000 (pp. 1-8).

Norton, L. S. (1990). Essay-writing: what really counts?. Higher Education, 20(4), 411-442.

Page, E. B. (1966). The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. The Journal of experimental education, 62(2), 127-142.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015, September). Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 431-439).

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). Automated essay scoring: A cross-disciplinary perspective. Routledge.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. Statistical Methodology, 6(2), 157-163.

An Adjusted BERT Architecture for The Automatic Essay Scoring Task

Ridha Hussein Chassab[a], Lailatul Qadri Zakaria[b]*, Sabrina Tiun[c]

[a,b,c]The Asean Natural Language Processing (ASLAN), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor Darul Ehsan, Malaysia.
* Email: lailatul.qadri@ukm.edu.my

Abstract

Automatic Essay Scoring (AES) is the process of identifying an automatic score for an essay answer. The state-of-the-art in AES task relies on word embedding techniques. One of the advanced embedding architecture that seems promising is the Bidirectional Encoder Representations from Transformers (BERT). Yet, such an architecture suffers from 'catastrophic forgetting' problem. This problem occurs because the gradients of fine-tuning BERT quickly forget significant information. In order to overcome such a limitation and to adapt the BERT architecture to be fit for the AES task, it is imperative to address an adequate adjustment on the learning rate. Therefore, this paper aims at proposing an adjusted BERT architecture based on unfreezing fine-tune mechanism in which the BERT architecture can adequately adopted for the AES task.

Keywords: Automatic Essay Scoring; Automatic Essay Grading; Bidirectional Encoder Representations from Transformers.

1. Introduction

Assessment is considered as the main component in the educational domain where student's abilities are being evaluated (Valenti, Neri, & Cucchiarelli, 2003). The last two decades showed a great investment in E-learning by utilizing the latest technologies. The dawn of 21 century has witnessed a great progress in terms of computerized assessment systems (Valenti et al. 2000). The majority of such systems were concentrating on multiple choice questions where the assessment would take the form of storing correct answers in a relational database and accommodating a simple matching mechanism to identify whether the student's answer is correct or not. Since the multiple-choice questions have a factual and exact answers, it was to implement an automatic assessment system to evaluate answers. Yet, another challenge has been arisen toward the computerized exams, such challenge is represented by the assessment of questions that require subjective answers. Automatic Essay Scoring (AES) is the process of identifying an automatic score for an essay answer. Subjective or essay answers require much more effort in order to articulate an automatic assessment for it.

The state-of-the-art in AES task relies Word2Vec and pretrained Glove architectures (Chen & Zhou, 2019; Hendre, Mukherjee, Preet, & Godse, 2020; Li, Chen, & Nie, 2020; Li et al., 2018; Liu et al., 2019; Wang, Liu, & Dong, 2018; Zhang & Litman, 2019). The main limitation behind such architectures lies in its inability to handle sentence-level embedding and they suffer from 'out-of-vocabulary' problem. This problem occurs when a term would have no embedding vector within the training (i.e., unseen). A solution for the aforementioned problems depicted by the Bidirectional Encoder Representations from Transformers (BERT) architecture where it has a fixed vocabulary size and a mechanism of rooting the terms. Although BERT showed remarkable performance for tasks like question-answering yet, it showed incompetency when tested for the AES task (Mayfield & Black, 2020; Rodriguez, Jafari, & Ormerod, 2019). The reason behind such miscarriage is due to a well-known limitation behind the BERT architecture which is 'catastrophic forgetting' (Rodriguez et al., 2019). This problem occurs because the gradients of fine-tuning BERT quickly forget significant information.

According to (Lehečka, Švec, Ircing, & Šmídl, 2020), there are plenty of parameters within the BERT architecture in which any task can be suited through tuning such parameters. One of the attempts to suite the BERT architecture for a particular task is through alter the learning rate. However, the way of altering the learning rate is very domain-specific issue where the specified task (i.e., AES) must be considered (Howard & Ruder, 2018). Therefore, an adjustment to the learning mechanism for the AES task is needed. This paper aims to propose an adjusted BERT architecture based on unfreezing fine-tune mechanism for AES task to overcome 'catastrophic forgetting' problem.

2. Related Work

Liu et al. (2019) proposed a multi-way attention architecture for AES task. The proposed architecture contains a transformer layer at first which process pre-trained Glove word embedding of student's answer and model's answer. Then, the following layer represents the multi-way attention where three self-attention vectors are represented for the student's answer, model's answer and their cross vector respectively. This will be followed with an aggregation layer where word's position vectors will be added. The final layer contains the regressor where the score of the essay is being predicted. For this purpose, the authors have used a real-word educational dataset of questions and answers. Result of accuracy was 88.9%.

Zhang & Litman (2019) proposed a deep learning architecture for AES task. The proposed architecture begins with pre-trained word embedding vectors brought from Glove and processed via Convolutional Neural Network (CNN) layer. Then, the resulted features will be processed via Long Short Term Memory (LSTM) in order to generate sentence embedding for each answer. The key distinguishes of this study lies in adding a co-attention layer that consider the similar sentences between student's answer and model's answer. Lastly, the final layer will give the score for each answer. Using the Automated Student Assessment Prize (ASAP) benchmark dataset, the proposed architecture produces an accuracy of 81.5%.

Kyle (2020) examined the lexical sophistication for evaluating second language writing proficiency (L2). The authors have used a corpus for English placement test (i.e., TOEFL). Using some lexical features such as word and n-gram overlapping along with a semantic approach of LSA, the authors have applied a simple regression in order to predict the score of the tested answers.

Li et al. (2020) have proposed a deep learning method for AES task where two architectures of CNN and LSTM are being employed. First, the authors have processed the words' vectors of each answer through the CNN architecture in order to get the sentence embedding. For this purpose, a pre-trained model of Glove word embedding has been used. In addition, the resulted sentence embedding from CNN have been furtherly processed via the LSTM architecture in order to get the score. Using the benchmark dataset of ASAP, the authors have shown an accuracy of 72.65%.

Tashu (2020) have proposed a deep learning architecture for AES task. The proposed architecture begins with word embedding vectors generated by Word2Vec and process via CNN layer in order to extract n-gram features. Lastly, a recurrent layer called Bidirectional Gated Recurrent Unit (BGRU) is being used to predict the score of the answer. Using the benchmark dataset of ASAP, the proposed architecture showed an accuracy of 86.5%.

The advancement of deep learning architecture led to the emergence of Transformers which yield a novel mechanism in learning. Such mechanism lies in the synchronized bidirectional learning. Such an architecture led to the emergence of Bidirectional Encoder Representations from Transformers (BERT) embedding. BERT has a fixed and indexed pretrained model of embedding where a vocabulary of 30,000 English terms is being stored. BERT has shown remarkable superior performance in text generation applications.

However, recently, Rodriguez et al. (2019) have utilized the BERT architecture for the AES task. Using ASAP dataset, BERT showed an accuracy of 74.75%. The authors have compared the BERT against the LSTM

and the comparison showed that LSTM is still a competitor where it achieved an accuracy of 74.63%. The authors have justified such a miscarriage of BERT regarding a problem known as 'catastrophic forgotten' where the BERT architecture would forget quickly what it had learnt previously. Similarly, Mayfield and Black (2020) have proposed a BERT architecture for the AES task. The authors have utilized the pretrained BERT embedding and then apply the fine-tune. Using ASAP dataset, results of accuracy showed an average of 64.6% achieved by the proposed BERT.

3. Proposed Adjusted BERT

BERT is an advanced deep neural network architecture that is based on a transformer which is intended to encode text and attempt to learn its deep linguistic context (Devlin, Chang, Lee, & Toutanova, 2018). BERT architecture consists of two main models including pretraining and fine-tuning as shown in Fig. 2. The pretraining contains a masked language model where some tokens within the text is being masked and the target is to predict these masks. In addition, the pretraining contains a sentence prediction where the sentences of a text document are being processed as input and the output is a binary classification of whether these sentences are consecutive or not. On the other hand, the fine-tuning model in BERT aims at accommodating specific task such as question answering, document classification or document ranking. In this study, the aim of fine-tuning is to predict the score of an answer therefore, it would be document ranking.

3.1. Unfreezing Adjustment

In fact, the fine-tuning part of BERT has a remarkable drawback of forgetting contextual information. An attempt to solve this problem has been depicted in the study of Howard and Ruder (2018) where an unfreezing mechanism is used to adjust latter hidden layers to fit a particular task. To understand the unfreezing mechanism, let assume multiple hidden layers within the fine-tuning architecture as shown in Fig. 1. The earliest layers would depict learning general features such as relationships between embedding vectors. However, the approach toward further hidden layers would require learning very specific characteristics of the particular task. Therefore, rather than using the fine-tuning BERT architecture as it is like the studies of Rodriguez et al. (2019) and Mayfield and Black (2020), it is necessary to examine a suitable adjustment. To this end, the learning rate values of the latter hidden layers' gradients will witness a gradual unfreezing. This can be seen as a gradual increment of learning rates within the latter hidden layers' gradients.

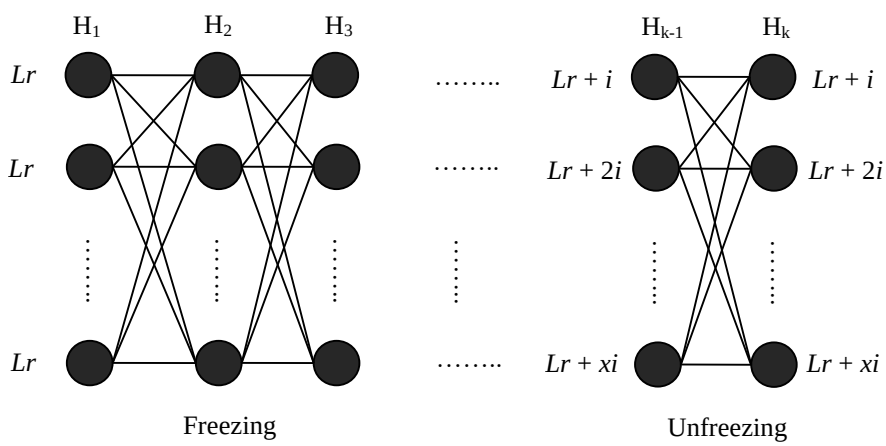


Fig. 1. Arbitrary hidden layers within BERT fine-tuning

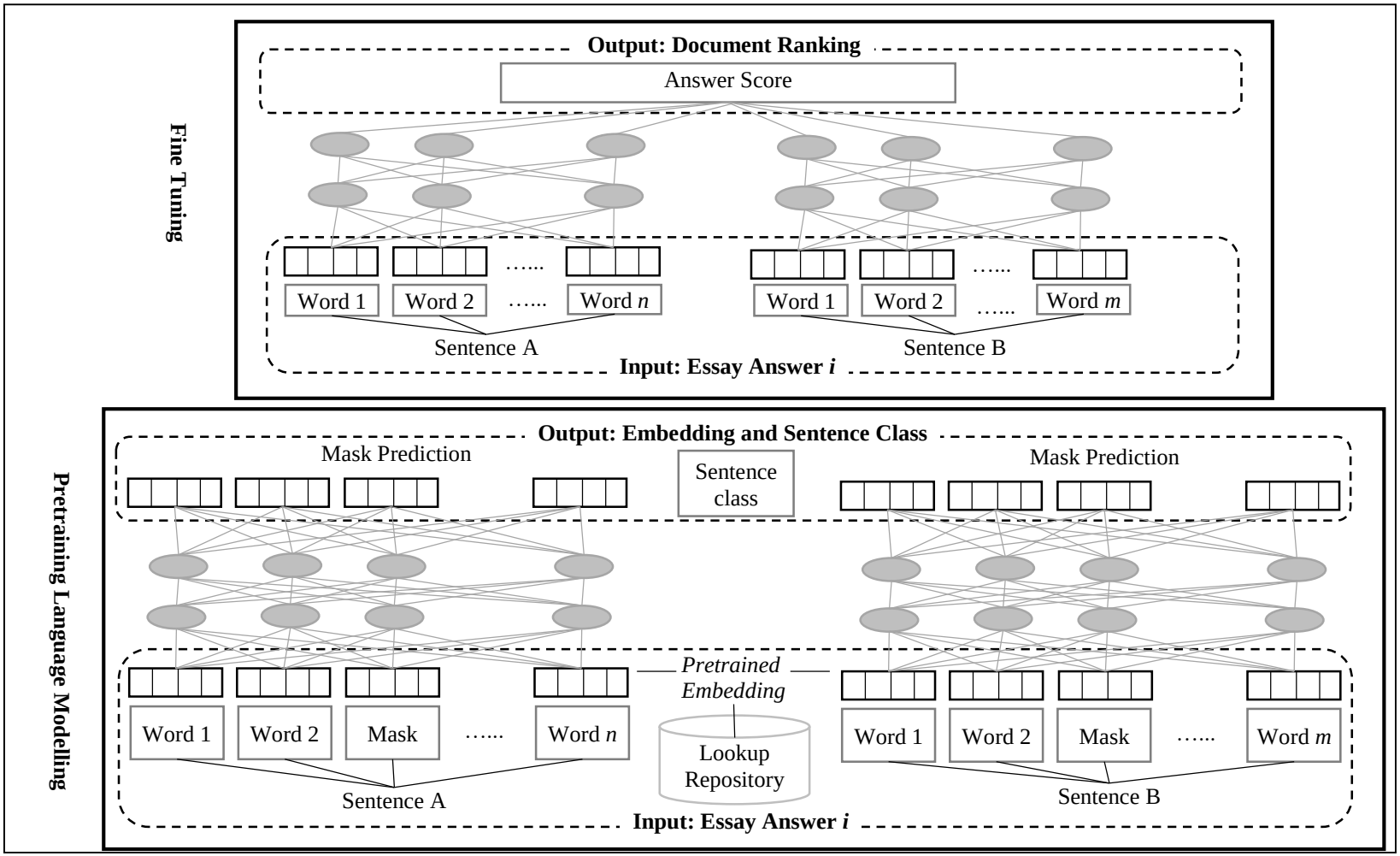


Fig. 2. The proposed BERT architecture for AES task

4. Conclusion

This paper has presented an adjusted BERT architecture for the AES task. Such an adjustment has been depicted by the unfreezing mechanism in which the learning rates of the latter hidden layers of the BERT's fine-tuning part are being gradually incremented. Such increment would contribute toward fit the learning model to the AES task. For future direction, the experimental results acquired by the proposed adjustment would have a valuable outcome in terms of examining the capabilities of BERT for the AES task.

Acknowledgements

This publication was supported by the Universiti Kebangsaan Malaysia (UKM) under GGP-2020-041.

References

Chen, Z., & Zhou, Y. (2019, 25-28 May 2019). *Research on Automatic Essay Scoring of Composition Based on CNN and OR.* Paper presented at the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hendre, M., Mukherjee, P., Preet, R., & Godse, M. (2020). Efficacy of Deep Neural Embeddings based Semantic Similarity in Automatic Essay Evaluation. *International Journal of Computing and Digital Systems, 9*, 1-11.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, 100467. doi:https://doi.org/10.1016/j.asw.2020.100467

Lehečka, J., Švec, J., Ircing, P., & Šmídl, L. (2020). *Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification*, Cham.

Li, X., Chen, M., & Nie, J.-Y. (2020). SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems, 210*, 106491. doi:https://doi.org/10.1016/j.knosys.2020.106491

Li, X., Chen, M., Nie, J., Liu, Z., Feng, Z., & Cai, Y. (2018). Coherence-Based Automated Essay Scoring Using Self-attention *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 386-397): Springer.

Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G. Y., & Liu, Z. (2019). *Automatic short answer grading via multiway attention networks.* Paper presented at the International Conference on Artificial Intelligence in Education.

Mayfield, E., & Black, A. W. (2020). *Should You Fine-Tune BERT for Automated Essay Scoring?* Paper presented at the Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications.

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and Automated Essay Scoring. *arXiv preprint arXiv:1909.09482*.

Tashu, T. M. (2020, 3-5 Feb. 2020). *Off-Topic Essay Detection Using C-BGRU Siamese*. Paper presented at the 2020 IEEE 14th International Conference on Semantic Computing (ICSC).

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research, 2*(1), 319-330.

Wang, Z., Liu, J., & Dong, R. (2018, 23-25 Nov. 2018). *Intelligent Auto-grading System.* Paper presented at the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS).

Zhang, H., & Litman, D. (2019). Co-attention based neural network for source-dependent essay scoring. *arXiv preprint arXiv:1908.01993*.

Informal Malay Language Twitter Corpus

Siti Noor Allia Noor Ariffin[a*], Sabrina Tiun[b]

[a,b] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, 43600, Malaysia
* Email: sitinoorallia@gmail.com

Abstract

In Malaysia, Twitter is a popular social media platform. This platform enables microblogging with a maximum of 280 characters per tweet. Users tweet almost everything that occurs during a single day. Due to its popularity, most Malaysians use Twitter daily, providing researchers and developers with abundant data on Malaysian users. This paper discusses how this study constructed a new Malay Twitter corpus and analyzed the data collected. The purpose of this paper is to compile tweets written in the informal Malay language. The data were extracted via Twitter's search function using relevant and related keywords associated with informal Malay language. The data was minimally pre-processed, as this study imposed several constraints on the collected tweets. The corpus data analysis reveals that most of the words in this corpus are informal, implying that Malaysians are most likely to write social media texts in informal Malay. This paper will benefit social media researchers and developers, particularly those with expertise in informal Malay and related fields.

Keywords: Informal Malay language; Malay Twitter corpus, Malay tweets

1. Introduction

Twitter is a social networking site that provides an online microblogging service that enables users of all backgrounds to send and read 280-character (Rosen & Ihara, 2017; Twitter, 2021) microblogs known as tweets. Tweets can be about anything, from jokes to current events to dinner plans (Britannica, 2020). According to Statista (2021), Twitter now has the most daily active users globally, surpassing the million-user mark in the fourth quarter of 2020 and remaining above that mark. Additionally, Statista (2021) reported that Malaysia was one of the top nations in the world in 2021, with approximately 3.35 million Twitter users. This analysis demonstrates that most Malaysians use Twitter by scrolling through feeds, retweeting, or saving content to retweet, which generates a wealth of data about Malaysian users. Twitter data can be easily collected using the application programming interface (API) provides by Twitter (Twitter, Inc., n.d.-b). It is, however, limited to the most recent seven days of data (Feizollah, Ainin, Anuar, Abdullah, & Hazim, 2019). To view tweets older than seven days, a premium account, which costs hundreds of dollars, is required (Twitter, Inc., n.d.-b). Furthermore, Twitter offers an Advanced Search feature that enables users to filter search results by date ranges, people, and more (Twitter Help Center, 2021). As a result, this study uses the Advanced Search feature to collect tweets containing informal Malay language and restricts the date range to February 2019.

The informal Malay language is a dialect of Malay used in everyday conversations by Malaysians. The language contains a large number of informal terms, including accent (or dialect) words, slang, titles (e.g. *hang*, *mek*), sounds (such as words written to express sounds like laughter, cat sounds, and knocking sounds), and mixed languages. The term "regional dialect or language" refers to a group of people who speak the language of a country state, resulting in word variation. By contrast, only a tiny percentage of the population understands slang. The term "mixed language" refers to the use of a foreign language in conjunction with Malay. When users write text on social media to provide reviews or opinions or tell a story, they frequently use everyday conversational language to project a friendly, casual, and easy-going image to other users.

Thus, this study aims to create a corpus of tweets written in the informal Malay language, encompassing various dialects, conversational slang languages, and mixed languages. According to previous research, the Malay Twitter corpus has existed since 2014. We discover the need for a corpus that incorporates dialect, informal, colloquial, and mixed-language while poring over the existing Malay Twitter corpus data. The tweets were gathered using Twitter's Advanced Search function (Supian, Razak, & Bakar, 2017; Ariffin & Tiun, 2018; Feizollah et al., 2019; Izazi & Tengku-Sepora, 2020) rather than the expensive, limited API (Feizollah et al., 2019). This data collection, however, is limited to the words contained in tweets. We purposefully ignored and omitted additional tweet features such as user information (full name & username), hashtags, URLs, and timestamps because we deemed them superfluous. We contributed to data collection and extraction by using various informal Malay languages from multiple Malaysia regions as keywords. Nevertheless, this dataset is not available for public use or future research because we still enforce the dataset's copyright.

The rest of this paper is structured in the following manner. Section 2 summarises pertinent works. Section 3 details the data collection and pre-processing procedures, while Section 4 present the corpus analysis and Section 5 summarises this work.

2. Literature Review

Social media has established itself as a valuable resource for researchers seeking to collect and curate massive amounts of data on a specific language or subject. Neunerdt and Zesch (2016) discovered that the primary characteristics of social media texts could be broadly classified as conversational language, dialogue styles, social media language, and informal writing. Conversational language is written language that transcribes spoken or everyday language. We were surprised to discover many slang terms and other colloquial expressions in a social media text, such as Twitter. According to Jamali (2018), Malaysian teenagers, on the other hand, are incredibly inventive when it comes to inventing new expressions and creative spelling. The manner of speaking (or dialogue styles) is also indicative of the writing style. For instance, the writer wishes to recount an event that occurred to the reader (or other users). Social media's language uses interaction signs such as emoticons, interaction words, leetspeak, word transformations, and conversational language. Simultaneously, informal writing contains errors such as spelling, abbreviation, sentence structure, and grammatical errors.

The Twitter application programming interface (API) is widely regarded as the de facto standard method for researchers and developers to extract data from Twitter. This API enables researchers to locate, retrieve, interact with, and create various resources, such as tweets, users, direct messages, lists, trends, media, and locations (Twitter, Inc., n.d.-a). Numerous previous studies have collected tweets from Twitter using this API. For instance, Maskat et al. (2020) retrieved and analyzed tweets about cyberbullying in Malaysia using the Twitter API. They analyze cyberbullying text and devise a method for automatically classifying tweets as "bully" or "not bully". Bakar, Rahmat, and Othman (2019) published a similar study in which they used the Twitter API to collect Malay tweets to conduct sentiment analysis and develop a polarity classification tool. In another work, Xu and Zhang (2018) analyze tweets about the #MH370 tragedy to develop a model of crisis information sharing based on sentiment, richness, authority, and relevance. They retrieved related tweets daily for 24 days using the Twitter API. In contrast to the previously mentioned works, we collected tweets in this study by configuring the Twitter Advanced Search feature (Twitter Help Center, 2021) to display only tweets containing the specified keyword within the specified date ranges (Supian et al., 2017; Ariffin & Tiun, 2018; Feizollah et al., 2019; Izazi & Tengku-Sepora, 2020). We chose this technique to avoid the API's monthly fee, ranging from $149 to $2,499 and its restrictions on data requests, most notably the number of requests (Feizollah et al., 2019).

Since 2010, indigenous researchers have created a slew of Malay corpora. The majority of early work on the Malay corpus (Don, 2010; Sidi et al., 2011; Mohamed, Omar, & Ab Aziz, 2011; Chung, 2011; Darwis, Abdullah, & Idris, 2012; Alshalabi, Tiun, Omar, & Albared, 2013; Bukhari, Anuar, Khazin, & Abdul, 2015;

Hoogervost, 2015; Hijazi et al., 2016; Yeo & Ting, 2017) concentrated on developing corpora containing formal Malay language content such as newspapers, speech texts, academy texts, and more. Nonetheless, as previously stated, the study and development of the Malay Twitter corpus (Arshi Saloot, Idris, Aw, & Thorleuchter, 2014; Omar et al., 2017; Anbananthen, Krishnan, Sayeed, & Muniapan, 2017; Ariffin & Tiun, 2018; Raja, Lay-Ki, & Su-Cheng, 2019) began in 2014, with the corpus consisting primarily of tweets written in informal Malay language. However, several previous studies' corpora (Omar et al., 2017; Raja et al., 2019) also included additional social media content, such as Facebook posts. Additionally, as mentioned earlier, we identify a need for a new Malay Twitter corpus that includes dialect, informal, colloquial, and mixed-language usage. According to our findings, the existing Malay Twitter corpus content does not possess the desired characteristics for accomplishing the study's objective. Thus, to achieve the purpose of this study, we retrieved tweets using previously identified keywords (Hoogervost, 2015; Hashim et al., 2016; Jamali, 2018) relevant and related to the informal Malay language. Furthermore, to ensure that we collected the correct and appropriate tweets, we used the findings of several previous studies as a guide for identifying the structure and other informal Malay languages (Kob, 2008; Hasrah & Aman, 2010; Sharum & Hamzah, 2011; Mansor, Mansor, & Rahim, 2013; Harun & Yusof, 2015; Jalaluddin, 2015; Jamil & Yusof, 2015; Sahril, 2016; Subet & Daud, 2016; Omar et al., 2017; Yeo & Ting, 2017; Choi & Chong, 2017; Jaafar, Aman, & Awal, 2017; Bakar & Mazzalan, 2018; Yusof, 2018; Wahab, 2018; Shafiee et al., 2019; Bakar & Tarmizi, 2019).

3. Methodology

This study aims to build a corpus of tweets written in the informal Malay language, including various dialects, conversational slang languages, and mixed languages. The methodology proposed entails data collection and pre-processing. As previously stated, the tweets were gathered using Twitter's Advanced Search feature. It searched using the provided keywords. Pre-processing is performed on the data by removing duplicate tweets from the corpus.

We manually collected data for this study, using keywords relevant and related to informal Malay language and limiting the date ranges to February 2019. The keywords were chosen after conducting a literature review on informal Malay language and structure. As mentioned previously, Twitter's standard method of collecting tweets is via their application programming interface (API), enabling developers and researchers to collect data. On the other hand, the API has numerous limitations, including a seven-day limit on tweets and a cap on the number of requests to the Twitter server (Feizollah et al., 2019). Hence, we chose to gather data using Twitter's Advanced Search feature manually, and the previously mentioned limitations became irrelevant.

Moreover, as previously stated, this study's data will be pre-processed by removing duplicates tweets from the corpus. This study used minimal data pre-processing because the collected data consisted solely of the words in the tweets and lacked additional tweet features such as user information (full name & username), hashtags, URLs, and timestamps. The pre-processing of the data begins with the identification and removal of duplicate tweets from the corpus. To begin, we sorted the data lexicographically ascending in order to identify any lines with repeated tweets. The lines containing repeated tweets, colloquially referred to as duplicated tweets, were then manually deleted from the corpus.

This study focuses exclusively on one criterion for tweet inclusion: tweets must be written in informal Malay. As explained earlier, informal Malay is a language rich in informal terms such as dialect, slang, titles, sounds, and mixed languages. Therefore, to ensure that the tweets we chose were appropriate and accurately reflected our study's objective, we used only keywords derived from previous research findings. On the contrary, this study currently does not have any exclusion criteria for tweets. We compiled a list of all tweets that were returned in response to the applied keywords. However, as mentioned earlier, we deliberately overlooked and omitted several additional tweet features. Other tweet features were deemed superfluous, as the study's

objective is to collect only text written in informal Malay. We disregard other characteristics to concentrate entirely on the textual characteristics and the frequency with which informal Malay was used in social media texts.

4. Analysis

According to our findings, the most frequently occurring words in this corpus are informal terms: *aku* (me/I; 887), *ni* (this; 548), *tu* (that; 537), *nak* (want; 514), *dia* (he/she/him; 389), *tak* (no; 374), *la* (358), *dah* (done; 281), *yg* (which; 274), and *nk* (want; 229). This data indicates that Malaysians are most likely to use informal Malay language when writing social media texts. Neunerdt and Zesch (2016) identify conversational language, dialogue styles, social media language, and informal writing as the primary characteristics of social media texts. Our findings indicate that the language and structure of our corpus conform to all the characteristics mentioned above. Thus, it demonstrates that, although our corpus is restricted to the words contained in tweets due to our deliberate omission of additional tweet features, our corpus still accomplishes our study's objective.

To aid researchers in better comprehending this corpus's fundamental properties, we provide several potentially useful statistics. The Twitter Advanced Search feature was used to retrieve 1,796 tweets written in informal Malay. The final dataset includes 38,714 tokens and 5,387 different word types. Additionally, the following are the values for fundamental n-grams, as well as their total number of tokens and types: unigrams (token: 37,382; types: 8,150), bigrams (token: 37,381; types: 32,280), trigrams (token: 37,380; types: 36,940), and 4-grams (token: 37,379; types: 37,207).

5. Conclusion

Overall, this work gathered tweets written in informal Malay language from Twitter. The data was pre-processed minimally to eliminate duplicate tweets. The analysis of the corpus data reveals that informal terms are the most frequently occurring words in this corpus. This analysis demonstrates that, despite the corpus data being restricted to the words contained in tweets without any additional tweet features, the corpus achieves the study's objective. Numerous aspects of this work can be enhanced in the future. For instance, the corpus size can be increased by extending the date ranges over which tweets are collected, and the keywords used to extract related and relevant tweets can be improvised by learning new keywords from other new research or by learning the variation for each word. This work is advantageous for those specializing in informal Malay and other related fields.

Acknowledgements

Universiti Kebangsaan Malaysia partially funds this research work under the research grant code: FRGS/1/2020/ICT02/UKM/02/1

References

Alshalabi, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the use of feature selection and machine learning methods in automatic malay text categorization. Procedia Technology, 11, 748-754.
Anbananthen, K. S. M., Krishnan, J. K., Sayeed, M. S., & Muniapan, P. (2017). Comparison of stochastic and rule-based POS tagging on Malay online text. American Journal of Applied Sciences, 14(9), 843-851.
Ariffin, S. N. A. N., & Tiun, S. (2018). Part-of-Speech Tagger for Malay Social Media Texts. GEMA

Online® Journal of Language Studies, 18(4).
Arshi Saloot, M., Idris, N., Aw, A., & Thorleuchter, D. (2014). Twitter corpus creation: The case of a Malay Chat-style-text Corpus (MCC). Digital Scholarship in the Humanities, 31(2), 227-243.
Bakar, M. S. A., & Mazzalan, A. M. (2018). Aliran Pertuturan Bahasa Rojak Dalam Kalangan Pengguna Facebook Di Malaysia. e-Academia Journal, 7(1).
Bakar, M. S. A., & Tarmizi, N. S. A. (2019). Penggunaan Dan Penerimaan Bahasa Slanga Dalam Novel Indie Di Malaysia. e-Academia Journal, 8(1).
Bakar, N. S. A. A., Rahmat, R. A., & Othman, U. F. (2019). Polarity classification tool for sentiment analysis in Malay language. IAES International Journal of Artificial Intelligence, 8(3), 259.
Britannica, T. Editors of Encyclopaedia (2020, November 18). Twitter. Encyclopedia Britannica. https://www.britannica.com/topic/Twitter
Bukhari, N. I. B. A., Anuar, A. F., Khazin, K. M., & Abdul, T. M. F. B. T. (2015). English-Malay code-mixing innovation in Facebook among Malaysian University students. Researchers World, 6(4), 1-10.
Choi, K. Y., & Chong, S. L. (2017). Campur aduk bahasa Melayu dan bahasa Cina. Journal of Modern Languages, 18(1), 56-66.
Chung, S. F. (2011). Uses of ter-in Malay: A corpus-based study. Journal of Pragmatics, 43(3), 799-813.
Darwis, S. A., Abdullah, R., & Idris, N. (2012). Exhaustive affix stripping and a Malay word register to solve stemming errors and ambiguity problem in Malay stemmers. Malaysian Journal of Computer Science, 25(4), 196-209.
Don, Z. M. (2010). Processing natural Malay texts: A data-driven approach. Trames, 14(1), 90-103.
Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., & Hazim, M. (2019). Halal products on Twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. IEEE Access, 7, 83354-83362.
Harun, K., & Yusof, M. (2015). Komunikasi bahasa Melayu-Jawa dalam media sosial. Jurnal Komunikasi: Malaysian Journal of Communication, 31(2).
Hashim, Nasihah & Mahmor, Noor & Ahmad, Ainal & Yahya, Maizatul Azura. (2016). BAHASA SLANGA DALAM KOMIK KANAK-KANAK (SLANG IN CHILDREN'S COMIC).
Hasrah, M. T., & Aman, R. (2010). Variasi Dialek Pahang: Keterpisahan Berasaskan Jaringan Sungai. Jurnal Melayu, 5.
Hijazi, M. H. A., Libin, L., Alfred, R., & Coenen, F. (2016, October). Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language. In 2016 2nd International Conference on Science in Information Technology (ICSITech) (pp. 356-361). IEEE.
Hoogervost, T. (2015). Malay youth language in West Malaysia. Youth Language in Indonesia and Malaysia. NUSA, 58, 25-49.
Izazi, Z. Z., & Tengku-Sepora, T. M. (2020). Slangs on Social Media: Variations among Malay Language Users on Twitter. Pertanika Journal of Social Sciences & Humanities, 28(1).
Jaafar, M. F., Aman, I., & Awal, N. M. (2017). Morfosintaksis Dialek Negeri Sembilan dan Dialek Minangkabau (Morphosyntax of Negeri Sembilan and Minangkabau Dialects). GEMA Online® Journal of Language Studies, 17(2).
Jalaluddin, N. H. (2015). Penyebaran Dialek Patani di Perak: Analisis Geolinguistik. Jurnal Antarabangsa Dunia Melayu, 8(2), 310-330.
Jamali, N. (2018). Fenomena Penggunaan Bahasa Slanga dalam Kalangan Remaja Felda di Gugusan Felda Taib Andak: Suatu Tinjauan Sosiolinguistik. Jurnal Wacana Sarjana, 2(3), 1-1.
Jamil, N. S., & Yusof, M. (2015). Analisis deiksis dialek Kedah. GEMA Online® Journal of Language Studies, 15(1).
Kob, M. A. C. (2008). Subklasifikasi Dialek Melayu Patani-Kelantan-Terengganu: Satu Analisis Kualitatif. Jurnal Melayu, 3.
Mansor, N. R., Mansor, N., & Rahim, N. A. (2013). Dialek Melayu Terengganu: Pendokumentasian dan

pengekalan warisan variasi bahasa tempatan. Jurnal Melayu, 10.
Maskat, R., Faizzuddin Zainal, M., Ismail, N., Ardi, N., Ahmad, A., & Daud, N. (2020, December). Automatic Labelling of Malay Cyberbullying Twitter Corpus using Combinations of Sentiment, Emotion and Toxicity Polarities. In 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence (pp. 1-6).
Mohamed, H., Omar, N., & Ab Aziz, M. J. (2011, June). Statistical malay part-of-speech (POS) tagger using Hidden Markov approach. In Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on (pp. 231-236). IEEE.
Neunerdt, M., & Zesch, U. D. I. T. (2016). Part-of-Speech tagging and detection of social media texts (pp. 1-123). RWTH Aachen University, Germany.
Omar, N., Hamsani, A. F., Abdullah, N. A. S., Abidin, S. Z. Z., & Alam, S. (2017). Construction of Malay Abbreviation Corpus Based on Social Media Data. Journal of Engineering and Applied Sciences, 12(3), 468-474.
Raja, R. A., Lay-Ki, S., & Su-Cheng, H. (2019). Exploring Edit Distance for Normalising Out-of-Vocabulary Malay Words on Social Media. In MATEC Web of Conferences (Vol. 255, p. 03001). EDP Sciences.
Rosen, A., & Ihara, I. (2017). Giving you more characters to express yourself. Retrieved from https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html
Sahril, S. (2016). PEMERTAHANAN BAHASA IBU MELALUI GRUP WhatsApp. Ranah: Jurnal Kajian Bahasa, 5(1), 43-52.
Shafiee, H., Mahamood, A. F., Manaf, A. R. A., Yaakob, T. K. S. T., Ramli, A. J., Mokhdzar, Z. A., ... & Ali, M. E. M. (2019). Pengaruh Bahasa Rojak Di Media Baharu Terhadap Bahasa Kebangsaan. International Journal, 4(15), 141-153.
Sharum, M. Y., & Hamzah, Z. A. Z. (2011). Golongan dan rumus kata gandaan berima. Jurnal Bahasa11, 27-47.
Sidi, F., Jabar, M. A., Selamat, M. H., Ghani, A. A. A., Sulaiman, M. N., & Baharom, S. (2011). Malay interrogative knowledge corpus. American Journal of Economics and Business Administration, 3(1), 171-176.
Statista (2021). Global Social Networks Ranked by Number of Users. Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
Subet, M. F., & Daud, M. Z. (2016). “Giler” Atau “Gile”: Slanga Kata Penguat.
Supian, M. N. A. A., Razak, F. A., & Bakar, S. A. (2017, April). Twitter communication during 2014 flood in Malaysia: Informational or emotional?. In AIP Conference Proceedings (Vol. 1830, No. 1, p. 020020). AIP Publishing LLC.
Twitter (2021). How to Tweet. Retrieved from https://help.twitter.com/en/using-twitter/how-to-tweet
Twitter Help Center. (2021, April 30). How to use advanced search – find Tweets, hashtags, and more. Twitter, Inc. https://help.twitter.com/en/using-twitter/twitter-advanced-search
Twitter, Inc. (n.d.-a). Getting Started with Our Platform. Twitter Developer. Retrieved May 21, 2021, from https://developer.twitter.com/en/docs/getting-started
Twitter, Inc. (n.d.-b). Twitter API Documentation. Twitter Developer. Retrieved May 18, 2021, from https://developer.twitter.com/en/docs/twitter-api
Wahab, K. A. (2018). Ciri dan Fungsi Komunikatif Bahasa Melayu Sabah dalam Media Sosial. Jurnal Komunikasi: Malaysian Journal of Communication, 34(4).
Xu, W. W., & Zhang, C. (2018). Sentiment, richness, authority, and relevance model of information sharing during social Crises—the case of# MH370 tweets. Computers in Human Behavior, 89, 199-206.
Yeo, D., & Ting, S. H. (2017). Netspeak features in Facebook communication of Malaysian university students. Journal of Advanced Research in Social and Behavioural Sciences, 6, 81-90.
Yusof, M. (2018). Trend ganti nama diri bahasa Melayu dalam konteks media sosial. Jurnal Komunikasi: Malaysian Journal of Communication, 34(2).

A Student Performance Model Towards Student Performance Prediction

Nor Samsiah Sani[a], Ahmad Fikri Mohamed Nafuri[b]*

[a,b]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
*Email: p102926@siswa.ukm.edu.my

Abstract

Student's performance prediction research on higher education commonly usesexamination grades and results as the main attribute.Contrastingly,in our research, the aim is todevelopa predictive model ofstudent's performance using student's data that has been enriched with data derived from students activities and employment status as the prediction label.This paper focuses on the early part of model development which is data preprocessing. In data preprocessing, several measures have been taken such as data integration, data cleaning, data transformation and correlation analysis using Spearman correlation. The results show that there are 15significant attributes with higher relation to the employment status.

Keywords: student performance; pre processing; prediction

1. Introduction

The ability to predict student'sacademic performance is very important in the field of education. The impact on academic performance comes from a variety of sources such as personal, social, psychological and environmental factors (Al-Hagery et al., 2020). Student's academic performance issue is highly debated, especially in tertiary education, because it has a direct impact on employability chances (Bhagavan et al., 2020). All parties in the field of education such as educational institutions, instructors, students and researchers can benefit by applying learning analytics on higher education data. For example, Avella et al. (2016)emphasised that the benefits includetargeted course offerings, curriculum development, student learning outcomes, personalized learning, improved instructor performance,post-educational employment opportunities and enhanced research in the field of education. Moreover, the results obtained such as significantly interesting samples, trends and even hidden information can help stakeholders in improving the process of teaching, exploration and description of phenomena occurring in the field of education (Osmanbegović & Suljić, 2012).

Therefore, the use of data mining methods in achieving the objectives of study is viewedas very appropriate because of its ability to handle large amount of data while identifyinghidden patterns and relationships (Bhardwaj, 2011). Hellas et al. (2018) stated in the study of predicting student performance that some methods that are often used by researchers can be categorized into several groups, namely classification (supervised learning), clustering (unsupervised learning), mining (finding frequent patterns and/or extraction of features) and statistics (correlation, regression and t-test). However, Zulkifli et al. (2019) suggested that predictive modelling for educational data in Malaysia are still lacking in terms of research number in order to yield a clear picture on student's academic performance in academic institutions.

Impact of different attributes on the student performance is widely reviewed and discussed in various researches. Researchers have included correlation method in analyzing the influencing factors in most of their works. Hutagaol & Suharjito (2019) studied the correlation between demographic and academic performance to predict student dropout and concluded that variables such as student's attendance, homework-grade, mid-test grade, and finals-test grade, total credit, GPA, student's area, parent's income, parent's education level, gender

and age as important features. Ajibade et al. (2020) predicted student academic performance using demographic, academic background, parents participation on learning process and behavioral features in a web based education system.

This paper aims to introduce the preprocessing part on all important features in the development of student's performance predictive model.

2. Methodology

One of the most important phase in building a student performance modelis data preprocessing. Steps involved in the preprocessing part were as follows:

Step 1: In data collection, the data used for this study was obtained and applied for from the Policy Planning and Research Division (BPPD), Ministry of Higher Education (MOHE). The raw data received is in comma separated values (CSV) format containing 248,568 public university graduates' data who had completed a bachelor's degree in 2015 to 2019. Table 1 lists the datasets included in this study.

Table 1. The list of datasets along with its information.

Dataset	Total of Attributes
Students	29
Activities	7
Awards	5
Industrial Training	4
MPP	3
Employment	5
Total	53

Step 2: In data data integration, six different datasets are combined into one dataset linked to student ID.

Step 3: In data cleaning, a data statistics review found that there are several records that have missing values.The attributes that contain some missing values are replaced with the most frequent value. Meanwhile Dewan Undangan Negeri (DUN), parliament and postal code attributes are removed entirely from datasets because the number of missing values were too high.Lastly, recurring attributes which listed similar records, are also removed from the dataset.

Besides that, the discretizationaccording to interval labels and conceptual labels has been made. Numerical attributes such as CGPA hasbeen discretizetofour labels namely first class, second class upper, second class lower and third class, accordingly.The purposes of the discretization are to handle noise, simplify the original data, improve data processing efficiency, generate more easily-described data representation and enhance the understanding of data mining results later on.

Step 4: In data transformation, attribute constructioncan help improve accuracy and understanding of the data. To leverage the date attributes available in the dataset, calculations and generation of new attributes such asage and duration of study have been performed.Feature engineering technique is then implemented in order to avoid duplication of student ID as theyhave several records in the activity and awards datasets.

Step 5: Correlation analysis. Before developing the predictive models, we conducted a statistical analysis using Spearman correlation to discern how significant the relationship between the study's class label (employment status) and the other variables is.

3. Results and Discussion

Student's data were collected and preprocessedas explained in step one to four. After the data were cleaned,Correlation analysis using Spearman correlation coefficient was conducted because it is suitable for dataset containingboth the continuous and discrete variables (Hussain et al., 2018). Every variable in the study received a correlation coefficient (r) after the Spearman correlation analysis, which represented the intensity and direction of the linear relationship between the tested pair of variables.

Table2. Correlation analysis and descriptive statistics for all students-related features.

Variables	r	P	Std. deviation	Mean
Jantina	0.030*	0.000	0.469	1.672
Perkahwinan	-0.043*	0.000	0.085	1.007
Negeri_Lahir	0.076*	0.000	1.398	2.861
Kediaman_Penginapan	0.016*	0.000	0.456	1.705
Kelas_B40	-0.011*	0.000	0.763	1.523
Sekolah	-0.019*	0.000	0.941	1.364
Universiti	-0.040*	0.000	0.656	2.210
Umur_Daftar	-0.009*	0.000	0.816	2.070
Kelayakan	-0.018*	0.000	0.806	2.047
Bidang_Pengajian	-0.052*	0.000	1.538	3.792
Tempoh_Pengajian	-0.018*	0.000	0.784	2.521
Tajaan	0.020*	0.000	0.782	2.347
Cgpa	0.043*	0.000	0.707	2.072
Keputusan_Li	0.049*	0.000	0.982	2.167
Bil_Aktiviti_Keseluruhan	-0.024*	0.000	6.573	2.945

Note: * Correlation is significant at the 0.05 and 0.01 levels, respectively (2-tailed)

The statistical results shown in Table 2 suggests that the correlation of *jantina, perkahwinan, negeri lahir*, *universiti, bidang pengajian, cgpa* and *keputusan LI* with the employment status ares lightly higher as compared to *kediaman penginapan, kelas B40, sekolah, umur daftar, kelayakan, tempoh pengajian, tajaan* and *bil aktiviti keseluruhan*. Moreover, Table 2 indicates that all variables were significant with respect to the employment status where P-values are less than 0.05.

This study concluded that the variables in our dataset are meaningful and can be used in subsequent studies to develop a prediction model of student performance.Machine learning algorithm such as random forest, decision tree, naïve bayes and k-nearest neighbour can be applied to our dataset because they are widely used and produce high accuracyaccording to prior research (Casuat & Festijo, 2020).

Acknowledgements

This publication was supported by the Universiti Kebangsaan Malaysia (UKM) under the Research University Grant (project code: GUP-2019-060).

References

Ajibade, S. S. M., Ahmad, N. B., & Shamsuddin, S. M. (2020). A data mining approach to predict academic performance of students using ensemble techniques. Advances in Intelligent Systems and Computing, 940, 749–760. https://doi.org/10.1007/978-3-030-16657-1_70

Al-Hagery, M. A., Alzaid, M. A., Alharbi, T. S., & Alhanaya, M. A. (2020). Data Mining Methods for Detecting the Most Significant Factors Affecting Students' Performance. International Journal of Information Technology and Computer Science, 12(5), 1–13. https://doi.org/10.5815/ijitcs.2020.05.01

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Journal of Asynchronous Learning Network, 20(2). https://doi.org/10.24059/olj.v20i2.790

Bhagavan, K. S., Thangakumar, J., & Subramanian, D. V. (2020). Predictive analysis of student academic performance and employability chances using HLVQ algorithm. Journal of Ambient Intelligence and Humanized Computing, 0123456789. https://doi.org/10.1007/s12652-019-01674-8

Bhardwaj, B. K. (2011). Data Mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, 9(4).

Casuat, C. D., & Festijo, E. D. (2020). Identifying the Most Predictive Attributes Among Employability Signals of Undergraduate Students. 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), February, 203–206. https://doi.org/10.1109/CSPA48992.2020.9068681

Hellas, A., Liao, S. N., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., & Messom, C. (2018). Predicting academic performance: a systematic literature review. Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE 2018 Companion, 175–199. https://doi.org/10.1145/3293881.3295783

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience, 2018, 1–21. https://doi.org/10.1155/2018/6347186

Hutagaol, N., & Suharjito, S. (2019). Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education. Advances in Science, Technology and Engineering Systems Journal, 4(4), 206–211. https://doi.org/10.25046/aj040425

Mayra, A., & Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in Ecuador. 2018 IEEE Global Engineering Education Conference (EDUCON), 1238–1242. https://doi.org/10.1109/EDUCON.2018.8363371

Osmanbegović, E., & Suljić, M. (2012). DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE. Economic Review: Journal of Economics and Business, 10(1), 3–12. http://hdl.handle.net/10419/193806

Zulkifli, F., Mohamed, Z., & Azmee, N. A. (2019). Systematic Research on Predictive Models on Students' Academic Performance in Higher Education. International Journal of Recent Technology and Engineering, 8(2S3), 357–363. https://doi.org/10.35940/ijrte.B1061.0782S319

A Dictionary Based Approach for Malay Language Sentiment Lexicon Generation

Azilawati Rozaimee[a*], Nazlia Omar[b], Sabrina Tiun[c] and NurSharmini Alexander[d]

[a]FIK, UniSZA, 22200, KampusBesut, Terengganu, Malaysia
[b,c]CAIT, UKM 43600, Bangi, Selangor, Malaysia
[d]MAMPU, 63000 Cyberjaya, Malaysia
*Email: azila@unisza.edu.my

Abstract

The sentiment lexicon plays a vital role in ensuring a successful sentiment analysis task. The most common approach to build one is via manual annotation. However, manual approach is labor-intensive, relatively slow, and requires much effort. This paper aims to automatically generate a Malay Language sentiment lexicon to overcome the limitation of the human-built sentiment lexicon. In this paper, we present the Malay Language Sentiment Lexicon generation algorithm based on a dictionary-based approach. This algorithm will utilize WordNet 3.0 and WordNet Bahasa resources, which are then mapped to build the new Malay Language Sentiment Lexicon. After the algorithm was run for five iterations from a pair of initial words, the generated sentiment lexicon produced a total of 61605 words with 25541 positive words and 36064 negative words. This shows that the proposed approach can generate a significant number of sentiment lexicons with reasonable accuracy for formal terms by utilizing dictionaries like WordNet 3.0 and Wordnet Bahasa.

*Keywords:*sentiment lexicon; Malay Language; dictionary-based; WordNet 3.0; Wordnet Bahasa

1. Introduction

Over the last few decades, most research on sentiment analysis has been done in English and other widely spoken languages such as Arabic and Chinese. English is a language that is available with many resources and tools for natural language processing (Alsaffar & Omar, 2015). Consequently, the need for more studies on sentiment analysis and construction of resources and tools for subjectivity and sentiment analysis in other low resources languages, such as Malay, is growing due to the increasing number of reviews in that language (Mate, 2016).One of the major challenges in sentiment analysis is the lack of resources. The primary problem for the development of sentiment analysis tools in Malay is almost none of the standard sentiment lexicon was developed (Nasharuddin et al., 2017). This paper will discuss the proposed method for the automatic sentiment lexicon generation for the Malay language. The remaining of this paper isas follows. Section II will discuss some related works on the dictionary-based approach. Next, section III describes the methodology and dataset used in this Malay sentiment lexicon generation development. Section IV will present the result and discussion of this study. Finally, the limitation and future works will be discussed in the conclusion section.

2. Related Works

Sentiment analysis is a highly active area of research that involves the computational study of opinions, evaluations, and reviews about products, services, and policies that are expressed in the written language, as well as the construction of sentiment corpora and dictionaries.There are two main approaches to perform sentiment analysis which are (i) machine learning approach; and (ii) lexicon-based approach. Most of the

sentiment analysis tools are built upon machine learning approach, which is supervised learning.However, this approach requires extensive training data to make sentiment analysis successful.

The alternative solution towards sentiment analysis with little or no training data is by having the lexicon-based approach. The lexicon-based approach applies classification weights associated with them, and the weight can be in binary polarities (positive/negative) or a numerical polarities format. Most of the research in unsupervised sentiment classification makes use of available lexical resources.One of the advantages of using a lexicon approach is that the lexicon can be built from a large corpus and then used in other applications where there may not be enough information to perform corpus-based approaches (Labille et al., 2017). This is in line with Nasharuddin et al., (2017) where lexicon-based approach does not require storing a large data corpus and training, so the whole process is much faster. The dictionary-based approach is one of the lexicon approaches which is a simple technique whereby it uses a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary (e.g., WordNet, Wordnet Bahasa). This approach starts with a manually collected seed set of positive and negative sentiment words, and then the expansion algorithm is iteratively executed to expand this set by searching in the dictionary for their synonyms and antonym and added to the seed list. After the expansion algorithm were run for a number of iterations, final list of latest seed set will be known as the sentiment lexicon. Other alternative methodsare hybrid and corpus-based approaches.

Several algorithms have been proposed to automatically generate sentiment lexicons using the dictionary-based approach for different languages across the world. Much work has been carried out for major languages such as English, Chinese, Spanish, and some other low resources languages such as the Amharic and Vietnamese language (Alemneh et al., 2020; Le et al., 2019). There is also prior work on sentiment lexicon that has been successfully developed for 136 major languages (Chen & Skiena, 2014). The constructed sentiment lexicon was done by appropriately propagating from seed words and resulting in high polarity agreement with published lexicons while achieving an acceptable lexical coverage.Another previous work also applies a dictionary-based algorithm to generate an Arabic sentiment lexicon that assigns sentiment scores to the words found in the Arabic WordNet (Mahyoub et al., 2014). This study works by linking the lexicon of AraMorph with SentiWordNetand shows that it can outperform state-of-the-art lexicon in terms of accuracy and F1-score. Other than that, Park & Kim (2016) propose a method to build a thesaurus lexicon for the Korean language. The dictionary-based approach uses three online dictionaries to collect thesauruses based on the seed words, and stores only co-occurrence words into the thesaurus lexicon to improve the reliability of the thesaurus lexicon. As for the Malay language, the algorithm for the dictionary-based approach was developed by Alexander & Omar (2017) and Darwich et al.(2016). The shortage of lexical resources that can assist in sentiment analysis task in the Malay language motivates the author to develop an algorithm that can automatically generate a standard Malay sentiment lexicon from the available Wordnet and Wordnet Bahasadictionary with lesser human intervention.

3. Methodology

The methodology used in this study was based on the dictionary-based sentiment lexicon generation approach. Figure 1 shows the model of Malay language sentiment lexicon generation. Sentiment lexicon generation phases start with a manually selected seed set. The seed set contains the list of the most important positive and negative words.The seed set will be expanded through the process of mapping words that exist in the seed set and match them withsynonyms and antonymsfound in the Wordnet 3.0 and Wordnet Bahasa using bootstrapping technique. This algorithmbeginsby matching the seed set with synsetid in Wordnet Bahasa to get the offset value. The expansion of the seed set works by matchingthe obtained offset valuetothe correspondingwordidto find the antonym and synonym of the positive and negative adjective which later will be added as the expanded lexicon. This expanded lexicon will become the seed set list for the next iteration.This process is iteratively done until no new words were found. In this work, this algorithm is repeated five times. The combination of the initial seed set, and expanded seed set generatedis finally called the sentiment lexicon,

which containsfinalized positive and negative adjective sentiment bearable words.

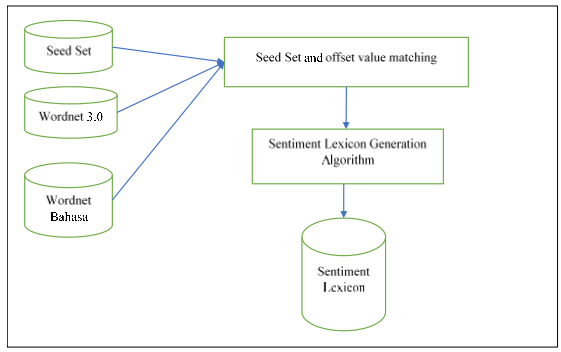


Fig. 1. Sentiment Lexicon Generation Model Based on Dictionary Approach

4. Results and Discussion

The result of the experiment is shown in Table 1.

Table 1. Total adjectives generated after five iterations.

Items	Total Number
Number of POS	25541
Number of NEG	36064
Total Words Generated	61605

The first iteration started with setting a pair of strong seed sets. The initial seed set was choose based on our benchmark study, which is “baik” for the positive seed set and “buruk” for the negative seed set. After the experiment was run for five iterations, the final iteration produceda total of 61605, with 36064 words that exist in negative polarity, meanwhile 25541 words with positive polarity. The result also shows that the final lexicon produced not only limited to a single word but also multiple terms such as "bersikapsabar”. It also can be seen that some of the terms were generated more than once. For example, in the positive list, the word 'suci' which means *clean* appears 45 times, and 'bangpak' which means *cruel* appear 18 times. Even though the same word appears multiple times, but it was rooted in different synsets.This result also shows that the synonymy generation process for positive sentiment words produced several same lemmas or words that duplicate with antonymy words generation in negative lexicon list. This also happened in synonymy generation for negative lexicon list, which also generated a few same lemmas in antonymy generation in the positive list.It can be seen here that the first iteration collects the related words in the synset mostly, but the number of sentiment words increases slowly in the remaining iterations. Due to this reason, the numbers of iterations were run at five (5) iterations only as being done in previous research (Alexander & Omar, 2017).

5. Conclusion and Future Works

In this paper, it was demonstrated that the sentiment lexicon generation algorithm for the Malay Languagebased ondictionary-based approach could lead to an expanded sentiment lexiconand promising results. However, this proposed work only contains adjectives, even though it is known that other Part-of-speech (POS) also carry the sentiment. Moreover, this work is only suitable to cater sentiment analysis task of formal

Malay due to the main resources for generating this Malaysentiment lexicon was only basedon formal Malay and does not include the informal Malay or slang.As far that we can see, user-generated data is greatly grown day by day and need to be analyzed.In future work, the enhancement of the proposed method will be explored to cater to the needs for generating sentiment bearable words from other classes of POS which are nouns, adverbs, and verbs. On top of that, an improved version of the algorithm that will make use of the corpus-based approach will be also considered so that it can fulfill the need for lexical resources on social media data. It is hoped that more significant coverage of the sentiment lexicon may enhance the quality of the sentiment analysis task.

Acknowledgement

The work was supported by Ministry of Higher Education for SLAB Scholars and the Universiti Kebangsaan Malaysia (GUP-2018-058).

References

Alemneh, G. N., Rauber, A., & Atnafu, S. (2020). Corpus based Amharic sentiment lexicon generation. 1–3. https://doi.org/10.18653/v1/2020.winlp-1.1

Alexander, N. S., & Omar, N. (2017). Generating a Malay Sentiment Lexicon Based on WordNet. Asia-Pacific Journal of Information Technology & Multimedia, 06(01), 127–141. https://doi.org/10.17576/apjitm-2017-0601-10

Alsaffar, A., & Omar, N. (2015). Study on feature selection and machine learning algorithms for Malay sentiment classification. Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014, 270–275. https://doi.org/10.1109/ICIMU.2014.7066643

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 2, 383–389. https://doi.org/10.3115/v1/p14-2063

Darwich, M., Mohd Noah, S. A., & Omar, N. (2016). Automatically Generating A Sentiment Lexicon For The Malay Language. Asia-Pacific Journal of Information Technology & Multimedia, 05(01), 49–59. https://doi.org/10.17576/apjitm-2016-0501-05

Kaity, M., & Balakrishnan, V. (2020). Sentiment lexicons and non-English languages: a survey. Knowledge and Information Systems, 62(12), 4445–4480. https://doi.org/10.1007/s10115-020-01497-6

Labille, K., Gauch, S., & Alfarhood, S. (2017). Creating Domain-Specific Sentiment Lexicons via Text Mining. Proc. Workshop Issues Sentiment Discovery Opinion Mining, 8, pages.

Le, C. C., Prasad, P. W. C., Alsadoon, A., Pham, L., & Elchouemi, A. (2019). Text classification: Naïve bayes classifier with sentiment Lexicon. IAENG International Journal of Computer Science, 46(2), 141–148.

Mahyoub, F. H. H., Siddiqui, M. A., & Dahab, M. Y. (2014). Building an Arabic Sentiment Lexicon Using Semi-supervised Learning. Journal of King Saud University - Computer and Information Sciences, 26(4), 417–424. https://doi.org/10.1016/j.jksuci.2014.06.003

Mate, C. (2016). Product Aspect Ranking using Sentiment Analysis : A Survey. 124–128.

Nasharuddin, N. A., Abdullah, M. T., & Azman, A. (2017). English and Malay Cross-lingual Sentiment Lexicon. Lecture Notes in Electrical Engineering, 2, 467–475. https://doi.org/10.1007/978-981-10-4154-9

Park, S., & Kim, Y. (2016). Building thesaurus lexicon using dictionary-based approach for sentiment classification. In 2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016 (pp. 39–44). https://doi.org/10.1109/SERA.2016.7516126

Analyzing Iraqi Dialects Unique Features for Dialect Identification

Ali Abdulraheem[a*], Lailatul Qadri Zakaria [b], Nazlia Omar[c]

[a,b,c] Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, University Kebangsaan Malaysia 43600 Bangi, Selangor Darul Ehsan, Malaysia

*Email: [a]aaj8068@gmail.com

Abstract

With the dramatic expansion of textual information, language identification has emerged as a task for analyzing such a huge amount of text. Dialect identification is a sub-task of language identification where a particular language and its sub-dialects are being addressed. This paper provides a series of features for improving the classification of Iraqi Arabic sub-dialects. It makes an effort to resolve the issue of sentence-level fine-grained Iraqi Arabic Dialects Identification of three distinct sub-dialects (Baghdadi, Maslawi, and Basrawi). Iraqi Arabic Dialects Recognition is a dynamic process in which other languages have common traits, such as having the same character and vocabulary. This paper aims to investigate an extensive space of features for identifying Iraqi Arabic sub-dialects by exploring a variety of feature extraction techniques such as (Special Character, POS features, Grammatical individual features, Case features, Gender features, Number features), as well as Machine learning-based models utilizing Multinomial Naive Bayes (MNB). However, this is the first preliminary analysis for Iraqi Arabic sub-dialects, which have not yet been interested in computational linguistics.

Keywords: Iraqi Arabic; Arabic morphology; Dialectal Arabic

1. Introduction

Arabic is one of the world's oldest languages it has been evolving over the decades. Arabic language can be classified into three categories: modern standard Arabic (MSA), classical Arabic (CA), and Arabic dialects (AD). MSA is formally used in official platforms including educational institutes, television broadcasts, and newspapers. CA is the language of the Holy Quran and Hadiths. It can also be viewed as the language of pre-Islamic poets. AD is the combination of different Arabic dialects spoken in different Arab countries. Such dialects have no written background, and they are formed by accommodating the varying degree of accents used in different cultures (Belkredim and Sebai 2009). Arab people use AD more than MSA in their everyday lives. AD is different from the CA and MSA in terms of morphology, phonology, lexicon, and syntax (Janet 2007). Different varieties of ADs are posing significant challenges for natural language processing tasks such as sentiment analysis, opinion mining, author profiling, and machine translation.

2. Related Work and Background

Arabic is known as a morphologically rich and complex language, which presents significant challenges for dialect identification. Arabic dialect identification is a crucial topic for most Arabic NLP research because of the diversity of the Arabic dialects. Some ADs in the same country shared features such as characters, vocabulary, and basic language set making, that amplifies the complexity of the dialect identification task.

Some studies have used different methods such as game-based theory (Alshutayri& Atwell 2018a; Osman et al. 2016) to automatically identify dialect in Arabic text. Bouamor et al. (2019), proposed a simple

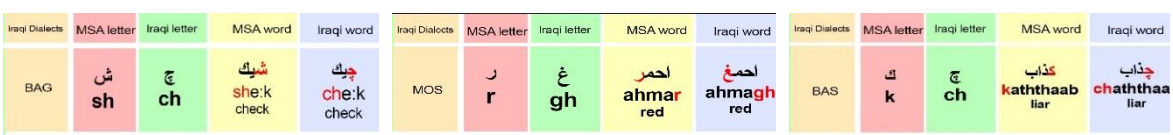
classification approach that only utilizes feature extraction word and character n-grams using Na¨ıve Bayes learning model. El-Haj et al. (2019) used grammatical, stylistic and Subtractive Bivalency Profiling features for dialect identification. Furthermore, several studies implemented a range of traditional word features such as N-gram, TF–IDF, PPS Tags, and POS and linguistic features with the machine learning techniques such as SVM, NB, and others (Ibrahim 2015, Kwaik et al. 2019; Obeid et al. 2019; Eltanbouly 2019) in order to identification dialects. Nonetheless, these approaches are facing limitations as each dialect has its own special morphology and phonology features, thus, discovering new features may facilitate dialect identification task.

3. Iraqi Dialect Overview

The Iraqi Arabic dialect is characterized by its diversity, often close to CA. According to Al-Rawi (2015), Iraqi dialects are divided into three dialects, including Baghdadi (BAG), Maslawi (MOS), and Basrawi (BAS) based on the location North, Centre, and South. This study encompassed a comprehensive background to illustrate the bounds of the Iraqi-Arabic dialects. It compares the phonological, morphological, orthographic, and lexical variations with Modern Standard Arabic. The Iraqi dialect contains 5 additional letters (che: چ, Pe: پ, Ve: ڤ, Gaf: گ, Ze: ژ) borrowed from the Indo-Iranian language in terms of pronunciation and writing, making the Iraqi Arabic alphabet 34 letters.

3.1. Phonology

Iraqi dialect consists of three sub-dialects which are Moslawi, Baghdadi, and Basrawi (Khoshaba 2006). In Mosul, Iraq's northernmost city, people speak the Moslawi dialect. In terms of pronunciation, Moslawi dialect keeps the letter"ق:q" as in MSA. While Baghdadi and Basrawi dialect use "ك:g" to replace the "ق:q". Southern dialect tends to change the "ك:k" in MSA to "چ:che". Fig 1 shows an example of the phonology differences between the letters in a BAG, MOS, and BAS dialects, respectively, with MSA.



(a) (b) (c)

Fig. 1. (a) difference between BAGand MSA,(b) difference between MOSand MSA,(c) difference between BASand MSA

The Iraqi dialects do not show the grammatical status of the term since they do not include diacritics such as fatha, damma, and kasra in the word. As a result, Iraqi phrases end with consonant characters rather than vowels. The Iraqi Arabic Dialect's past tense is influenced by ancient Mesopotamian languages and Assyrian dialects of Iraq's original inhabitants, and it begins with a consonant cluster (Khoshaba 2006). In Iraqi dialect, the future tense prefix is " راح , raħ, went". This " راح , raħ " it is inserted before the present tense verb to denote the future, similar to the English word "will," while in MSA the prefix is "س, s" or "سوف,swf, will " to imply the future tense. The prefix "دا /daʔ" or " جاي /gˇaˉyˉ/" is applied to the beginning of the verb to indicate the present tense, although the MSA does not have such thing (Khoshaba 2006). There is a common relative pronoun "اللي /aʔlly/" in Iraqi Arabic dialects, extracted from an MSA pronoun "الذي, aladhi, which" is used for singular masculine (Khoshaba 2006).

3.2. Lexicon

The Iraqi dialects are rich in lexicon, as many words are borrowed from the Mesopotamia civilization and languages from other neighboring countries, such as Turkish, Persian, Assyrian, and other Arabic dialects. for example, the word "روزنامه, rwznAmh /calendar " barrowed from Persian while the word "خاشوكهʾašwkh/spoon " from Turkish.

4. The Development of Iraqi Corpus

Iraqi dialects have increasingly been used in Social media which provide a good source of data for NLP tasks. We have developed an annotated morphological Iraqi corpus called Al-Rafidayn which contains 3,000 sentences. To develop a morphology corpus in Iraqi dialects, four stages have been adopted. In the first stage, annotation labeling is done by online web survey. Arabic Buckwalter transliteration and Lemma are used to recognize the Iraqi Indo-Iranian letters in the second stage. The third stage utilized morphological online tools to annotate grammatical properties. Finally, the fourth stage is an agreement which was reached with annotators to correct the error and verify the quality of the previous stages.

5. Methodology

This work aims to identify a set of features to improve Iraqi dialects classification. The study has adopted a variety of feature extraction and machine learning-based models using Multinomial Naive Bayes (MNB) in terms of training and testing.

5.1. Features Extraction

Extracting a set of discriminative features from the data helps in distinguishing the different classes. This study aims to extract specific Iraqi morphological and lexical handcrafted features to distinguish the Iraqi sub-dialects from the Iraqi annotated morphological corpus. This corpus is developed to include the inflection and diacritics due to the Iraqi dialects that use them as characters. For example, the word "لكِ, laki" which means "is yours", the diacritic kasra is replaced by the letter "ya, ي" to become "الكي, laky" in the MOS dialect to express the feminine pronoun, while in BAG and BAS they tend to onvert the letter "kaf and kasra" to the letter "Jim,ج" to becomes "الج, lağ". This may aid the feature extraction process to utilize a variety of linguistically motivated feature sets, namely morphological content. The features considered are shown in Table 1.

Table 1. Feature extracted for Iraqi Arabic Dialect

Features	Description
Special Character	The Special 5 Iraqi letters (Indo-Iranian letter) along with possible derivational inflections
POS features	nouns, Number of words, proper nouns, adjectives, adverbs, Number of Pronouns, verbs, particles, prepositions, abbreviations, punctuation, conjunctions, interjections, foreign letters.
Case features	nominative, accusative, and genitive.
Gender features	Feminine and Masculine.
Number features	singular words, plural words, and dual words.
Grammatical person features	1st person, 2nd person, 3rd person.

5.2. Feature selection

Feature selection is an optimization technique that narrows down the feature space by selecting a subset of the original set's most important features. In this work, the Random Forest algorithm is used to select the top features. It is an ensemble learning algorithm based on combining a number of de-correlated decision trees in which the tree-based structure is naturally used to rank the features.

5.3. Classification

This study performs multi-class classification in the experiments. In particular, Multinomial Naive Bayes (MNB), is a variation of Naive Bayes that estimates the conditional probability of a token given its class as the relative frequency of the token *t* in all documents to class *c*. MNB has proven to be suitable for classification tasks with discrete features (e.g. Word or character counts or representation for text classification) (Manning et al. 2008).

6. Conclusion

This study aims to identify the Iraqi Arabic dialects. To achieve this goal, an annotated morphosyntactic Iraqi dialects corpus includes three main dialects in Iraq (BAG, MOS, and BAS) has been created. Then, this corpus was used to train the proposed approach to extract features along with an MNB to identify the sub-dialects in the Iraqi dialect. For future directions, carrying out the experiments and analyzing the obtained results would be our next interest for determining the best subset of features.

Acknowledgements

This publication was supported by the Universiti Kebangsaan Malaysia (UKM) under GGP-2020-041.

References

Alshutayri, A. & Atwell, E. 2018a. Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers (May). Retrieved from http://eprints.whiterose.ac.uk/128607/
Alshutayri, A. & Atwell, E. 2018b. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. OSACT 3 Proceedings. LREC.
Bouamor, H., Hassan, S. & Habash, N. 2019. The MADAR shared task on Arabic fine-grained dialect identification. Proceedings of the Fourth Arabic Natural Language Processing Workshop, hlm. 199–207.
El-Haj, M., Rayson, P. & Aboelezz, M. 2019. Arabic dialect identification in the context of bivalency and code-switching. LREC 2018 - 11th International Conference on Language Resources and Evaluation 3622–3627.
Eltanbouly, S., Bashendy, M. & Elsayed, T. 2019. Simple But Not Naïve: Fine-Grained Arabic Dialect Identification Using Only N-Grams 214–218. doi:10.18653/v1/w19-4624
Ibrahim, H.S., Abdou, S.M. and Gheith, M., 2015. Sentiment analysis for modern standard Arabic and colloquial. arXiv preprint arXiv:1505.03105.
Khoshaba, M. P. 2006. Iraqi dialect versus standard Arabic. Medius Corporation.
Kwaik, K. A., Saad, M., Chatzikyriakidis, S. & Dobnik, S. 2019. Shami: A corpus of levantine Arabic dialects. LREC 2018 - 11th International Conference on Language Resources and Evaluation 3645–3652.

Manning, C. D., Raghavan, P. & Schütze, H. 2008. Text classification and naive bayes. Introduction to information retrieval 1(6).
Obeid, O., Salameh, M., Bouamor, H. & Habash, N. 2019. ADIDA : Automatic Dialect Identification for Arabic (iii): 6–11.
Osman, M., Sabty, C., Sharaf, N. & Abdennadher, S. 2016. Building a corpus for Arabic dialects using games with a purpose. Proceedings - 1st International Conference on Arabic Computational Linguistics: Advances in Arabic Computational Linguistics, ACLing 2015 21–25. doi:10.1109/ACLing.2015.10
Watson, J. C. E. 2007. The phonology and morphology of Arabic. Oxford University Press on Demand.

Security Assessment for Education Websites in Saudi Arabia

Almirabi Anas Anwar M[a], Mohd Zamri Murah[b]*

[a,b] Pusat Keselamatan Siber, Universiti Kebangsaan Malaysia
* Email: zamri@ukm.edu.my

Abstract

Many educational institutions use educational websites to improve teaching and learning. How-ever, these educational websites are open to cyberattacks. In this paper, we proposed a frameworkto access the cyber readiness of educational websites. We used education websites in Saudi Arabia as a case study. The framework consists of four phases: Reconnaissance, Enumeration, Scanning, Vulnerability Assessment, and Content Analysis. The reconnaissance phase uses OSINT technology, Enumeration and Scanning uses Nmap, Vulnerability Assessment using automatedscanning tools, and Content Analysis uses SSL tools. In our case study, we evaluated 12 Saudi Arabia educational websites. Our result indicated that cyber readiness for the 12 websites varies.We found many cybersecurity issues among the websites, such as outdated operating systems, unnecessary open ports, improper running services, a high number of web vulnerabilities, and low-grade SSL implementation. These issues, if not remedied, would provide a high probabilityof successful cyber attacks from hackers.

Keywords: web security, security assessment, penetration testing

1. Introduction

As more educational institutions seek to offer online learning and services, education web- sites have grown increasingly important (Mburano & Si, 2018). These websites, on the other hand, are attractive targets for cyberattacks for a variety of reasons. To begin with, distinguishingbetween a legitimate and malicious user is difficult. When a user interacts with a website, data is exchanged, and determining malicious data or exchange can be difficult. Second, web applications have grown in complexity and become more vulnerable to security flaws. Hackers couldtake advantage of these flaws to gain access to the system. Thirdly, design flaws, incorrect configuration, and a lack of updates can all lead to vulnerabilities. Fourth, educational institutions are frequently targeted by hackers because they are easily exploitable and contain a wealth of valu-able information. Data leakage, loss of privacy, financial effect, and loss of consumer trust are allconsequences of cyberattacks. For these reasons, websites should be prepared for cyberattacks with a high level of cybersecurity readiness (Shah & Mehtre, 2014).

This study presents a framework for assessing cyber readiness for educational websites. As a case study, we looked at educational websites in Saudi Arabia. We could detect threats and vulnerabilities using the framework and make recommendations to increase the cyber resilienceof websites.

In Saudi Arabia, there has been an increase in demand for online educational websites. EverySaudi citizen has free access to public education from primary school to college. Educationis Saudi Arabia's second-largest government expenditure, accounting for 8.8% of the country'sgross domestic product (Alotaibi, 2013). However, there is currently no study on the level of cyber readiness of educational websites in Saudi Arabia.

2. Methodology

The cyber readiness framework consists of 4 phases as shown in Figure 1; Reconnaissance, Enumeration and Scanning, Vulnerability Assessment, and Content Analysis. In the first phase,we used OSINT tools such as *shodan.io,zoomeye.io* and *sublist3r* to discover websites from do-main *edu.sa*. The results would provide a list of all hosts from that particular domain, includingIoT devices, websites, routers, and networks.

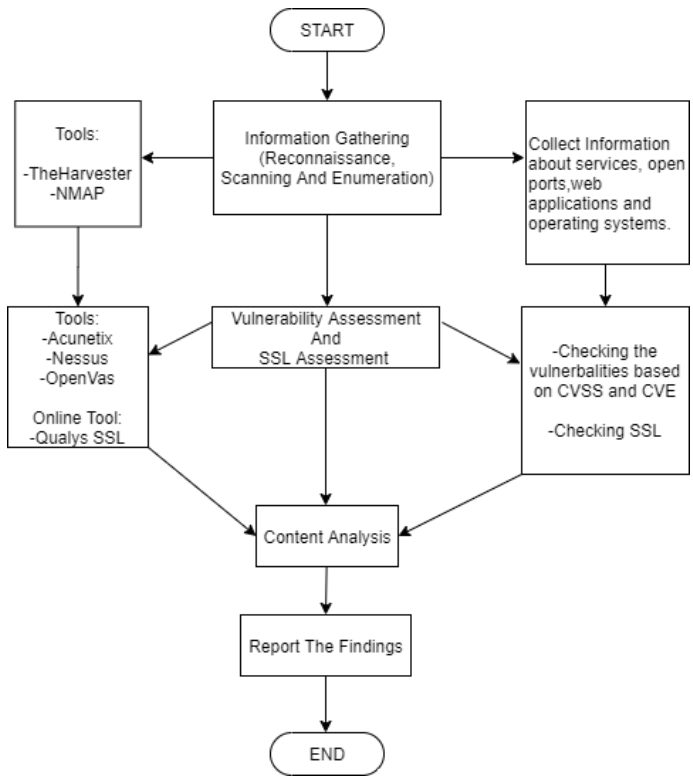


Fig.1. Assessment of cyber readiness framework. The framework consists of four phases: Reconnaissance, Enumeration and Scanning, Vulnerability Assessment, Content Analysis

In the second phase, we used *Nmap* to enumerate and scan the websites. Using *Nmap*, we would get information about OS, software versions, open ports, running services, and server type. This information would provide imexploit exploited portant attack vectors for hackers. Forinstance, if the website used Microsoft-IIS 7.5, the hackers could specific exploits such as login exploits or DDoS exploits for Microsoft-IIS 7.5 to compromise the website. Open ports could be vulnerable to buffer overflows or remote exploits. Running services indicate services that could be comprised, such as RDP and SSH. For example, an Eternal Blue exploited RDP vulnerability togain access into a system. Servers type would indicate whether the websites are running on Unix or Windows. There are different exploits for Unix-based systems and Window-based systems.

In the third phase, we used automated web vulnerability scanners *OpenVAS* (Rahalkar, 2019), *Nessus* (Chauhan, 2018), and *Acunetix* (Erturk & Rajan, 2017) to scan for vulnerabilities. The useof automated scanners is controversial. These scanners are automated and have comprehensive rules to test the websites. However, this scanning could hang the websites, causes network bot- tlenecks, and time consuming (Mburano & Si, 2018). These scanners produced results that varyfrom one another because each scanner algorithm for detecting and

identifying vulnerabilities differs (Alsaleh et al., 2017). Thus, the results from the scanners need to be manually verified. The servers typically log the scanning processes. Scanning usually would trigger the IDS to block the source IPs of the scanners. The phase also takes a long time because the scanners will ex- haust all rules to scan the websites. There are currently efforts to build scanners that use artificialintelligence rules to scan and save time.

In the fourth phase, we evaluated the SSL implementation using *Qualys SSL Labs*. Each website will be given a letter grade A to C based on their level of SSL implementation. The best grades are A and A+, indicating the websites have an excellent SSL implementation and certificate. We also manually look for security policies on the websites. A good website would have a security policy on how they handle privacy and confidential customer data.

3. Results and Discussion

In the first phase, we obtained 3,676 educational websites from the *edu.sa* domain. We began by choosing 29 websites. We looked over these websites to see if they were appropriate forour research. We didn't include sensitive government education websites or those that weren't updated on a regular basis. Finally, we decided to focus our case study on only 12 websites. We anonymised them to protect their privacy.

In the second phase, we found several websites were running Windows IIS 8.5 Server, a server released in 1995. This server's support was extended till 2020. Because the servers no longer gotsecurity fixes, websites that employed software that was not adequately supported were vulnerableto cyberattacks. Unnecessary services and open ports are running on some websites, which couldbe exploited by buffer overflow attacks or remote exploitation attacks. A few websites disclosetheir software version, providing attackers with even another attack channel.

In the third phase, the vulnerability scanners assign CVSS (Common Vulnerability Scoring System) scores to discovered vulnerabilities and use those scores to divide those vulnerabilitiesinto four categories: high (H), medium (M), low (L), and informational (INF) (I). The severity of a vulnerability is reflected in the vulnerability classification. For instance, a high vulnerabilityrating would indicate a vulnerability that would have a severe impact on the website, such as dataloss, unauthorized login access, or data breach.

We observed that *Acunetix* gave notable different results from *Nessus*, and *OpenVAS* was un-able to detect vulnerabilities in many websites. *Nessus* found 7 high vulnerabilities for w10, 1 for w11, and 2 for w12. *Acunetix* found 22 high vulnerabilities in w1, 851 in w11, 22 in w12 and 4 in w8. The high vulnerabilities were security issues that needed to be remedied. Medium vulnerabilities and informal vulnerabilities are acceptable risks that can be ignored. In the fourthphase, the result indicated w1, w2, w3, w4, w5, w6, w10, w12 as grade B, w7 as grade A, w8 as grade A+, w10 as grade F, and w11 receives no grade. We could conclude that 8 websites haveexemplary SSL implementation, 2 websites excellent SSL implementation. Websites that handlesensitive data should have SSL implementation of grade A or above.

Table 1: Vulnerabilities count based on *OpenVAS*, *Nessus* and *Acunetix*. The label H indicate High, M (Medium), L (Low) and I (informational). For instance, at host *w1*, *OpenVas* didn't find any vulnerabilities while *Nessus* found 30 vulnerabilities and *Acunetix* found 122 vulnerabilities

websites	OpenVAS	Nessus	Acunetix
w1	-	M(2) L(1) I(27)	H(22) M(58) L(11) I(31)
w2	-	M(2) I(20)	L(1) I(2)
w3	-	I(2)	M(843) L(2) I(233)
w4	-	I(14)	M(3) L(3) I(3)
w5	H(1) I(66)	I(72)	M(435) L(55) I(703)
w6	H(1) I(66)	I(46)	M(646) L(45) I(23)

w7	-	M(1) I(23)	M(1) I(1)
w8	-	I(8)	H(4) M(2) L(4) I(7)
w9	-	M(1) L(1) I(30)	M(18) L(11) I(32)
w10	-	H(7) M(6) I(17)	I(1)
w11	-	H(1) I(25)	H(851) M(3016) L(467) I(504)
w12	-	H(2) M(1) I(27)	H(22) M(58) L(11) I(31)

4. Conclusion

In summary, the cyber security readiness of Saudi Arabia's 12 educational websites varies. A number of websites are still using outdated operating systems, and as a result they are more vulnerable to cyberattacks. Sites which are not only open, but have open ports could be utilised for remote executions and buffer overflows. Those websites that have higher levels of vulnerabilityare prone to cyberattacks. A measure of the overall severity of the vulnerabilities dictates whichcyber issues need to be corrected and which can be ignored. The proper installation of SSL is absolutely essential for websites that handle sensitive data. Using the suggested framework, wecan use this to evaluate the cyber readiness of other educational websites. The defence plans outlined in these assessments will prove to be invaluable in the event of future cyberattacks.

References

Alotaibi, M.B. 2013. Assessing the usability of university websites in saudi arabia: A heuristic evaluation approach. *201310th International Conference on Information Technology: New Generations*. IEEE.

Alsaleh, M., Alomar, N., Alshreef, M., Alarifi, A. & Al-Salman, A. 2017. Performance-based comparative assessmentof open source web vulnerability scanners. *Security and Communication Networks*.

Chauhan, A.S. 2018. *Practical Network Scanning: Capture network vulnerabilities using standard tools such as Nmapand Nessus*. Packt Publishing Ltd.

Erturk, E. & Rajan, A. 2017. Web vulnerability scanners: A case study. *arXiv preprint arXiv:1706.08017*.

Mburano, B. & Si, W. 2018. Evaluation of web vulnerability scanners based on OWASP benchmark. *2018 26th International Conference on Systems Engineering (ICSEng)*. IEEE.

Rahalkar, S. 2019. Openvas. *Quick Start Guide to Penetration Testing*, pp. 47–71. Springer.

Shah, S. & Mehtre, B.M. 2014. An overview of vulnerability assessment and penetration testing techniques. *Journal ofComputer Virology and Hacking Techniques* 11(1): 27–49.

Aplikasi Mudah Alih Augmentasi Realiti Pembelajaran Kimia Berasaskan Soalan

Mobile Augmented Reality Application based on Questions for Learning Chemistry

Nur Atiqah Najibah Shamsudin[a] dan Nazatul Aini Abd Majid[b*]

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, UKM Bangi 43600, Selangor.
*Email: nazatulaini@ukm.edu.my

Abstrak

Augmentasi Realiti (AR) merupakan teknologi yang semakin menonjol dalam era revolusi industri keempat kerana keupayaannya untuk menggabungkan dunia realiti dengan dunia maya. Namun, dalam bidang pendidikan, gabungan di antara dua dunia ini perlulah dalam bentuk yang dapat meningkatkan lagi minat dan penglibatan pelajar dalam pembelajaran terutama dalam aspek kognitif. Untuk pembelajaran subjek kimia diperingkat sekolah menengah, minat pelajar didapati kurang kerana kesukaran untuk mengvisualkan gambaran objek atom bagi setiap elemen berserta cirinya ketika mempelajari konsep jadual berkala. Kajian ini mereka bentuk dan membangun aplikasi mudah alih AR berasaskan soalan untuk pelajar mengasah kemahiran kognitif mereka ketika menggunakan aplikasi ini dalam pembelajaran jadual berkala. Skop aplikasi ini merangkumi paparan animasi atom dalam bentuk model 3D untuk unsur bagi kumpulan 17 dalam jadual berkala. Aplikasi ini dibangunkan menggunakan perisian Unity 3D, Vuforia, Autodesk 3Ds Max, Audacity dan Adobe Photoshop berpandukan model pembangunan ADDIE. Aplikasi mudah alih ini telah dinilai oleh pelajar sekolah melalui kajian kes dan didapati aplikasi ini berjaya menarik minat dan penglibatan pelajar dalam mempelajari topik kimia. Kajian ini menunjukkan aplikasi AR yang mendorong pelajar menjawab soalan berdasarkan gabungan bahan pembelajaran realiti dan bahan digital berupaya meningkatkan usaha pelajar dari segi mental untuk mempelajari sesuatu topik.

Kata kunci: Augmentasi realiti; Jadual berkala; Kimia; STEM

Abstract

Augmented Reality (AR) is an increasingly prominent technology in the era of the fourth industrial revolution because of its ability to combine the world of reality with the virtual world. However, in the field of education, the combination between these two worlds must be in a form that can further increase students' interest and involvement in learning, especially in the cognitive aspect. For the study of chemistry subjects at the secondary school level, students' interest was found to be less due to the difficulty of visualizing the atomic object representation for each element and its characteristics when studying the concept of the periodic table. This study designed and developed a question -based AR mobile app for students to improve their cognitive skills while using this app in periodic table learning. The scope of the application includes the display of atomic animations in the form of 3D models for the elements of group 17 in the periodic table. The application was developed using Unity 3D, Vuforia, Autodesk 3Ds Max, Audacity and Adobe Photoshop software based on the ADDIE development model. This mobile application has been evaluated by school students through case studies and it was found that this application has succeeded in attracting the interest and involvement of students in learning chemistry topics. This study shows that AR applications that encourage students to answer questions based on a combination of reality learning materials and digital materials are able to increase students' mental efforts to learn a topic.

Keywords: Augmented reality; Periodic Table; Chemistry; STEM

1. Pengenalan

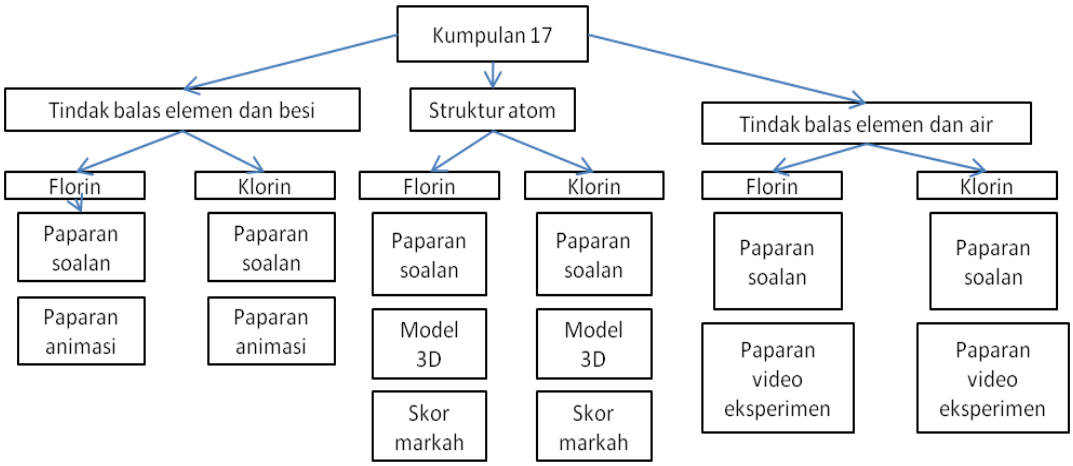
AR adalah satu teknologi yang membolehkan objek maya di dalam bentuk 2 Dimensi (2D) dan 3 Dimensi (3D) digabung ke dalam persekitaran nyata. Persekitaran pada dunia nyata dijejak oleh aplikasi AR berdasarkan penanda yang telah ditetapkan. Bahan digital berbentuk Objek 3D ini muncul di atas skrin telefon apabila kamera telefon dapat menjejak penanda atau lokasi yang telah direkodkan di dalam pangkalan data sistem (Lee 2012). Teknologi AR wujud apabila ianya divariasi daripada teknologi realiti maya iaitu realiti maya hanya fokus kepada dunia maya sahaja, manakala AR menggabungkan dunia nyata dan dunia realiti. Oleh kerana AR berkemampuan untuk meningkatkan persepsi pengguna dan membenarkan pengguna untuk berinteraksi dengan maklumat atau objek maya di ruang nyata, pendekatan baru ini dilihat menarik dan efisien untuk diimplementasi dalam proses pembelajaran dan pendidikan (Billinghurst & Duenser 2012).

Terdapat tiga ciri utama di dalam teknologi AR iaitu menggabungkan dunia nyata dan maya, berinteraksi dengan pengguna dalam waktu nyata dan diolah di dalam bentuk permodelan 3D. Namun, tidak semua jenis sukatan pelajaran sesuai untuk menggunakan teknologi ini. Pembangun harus mempertimbangkan keperluan pengguna dan keberkesanan hasil akhir. Kaedah pembelajaran yang sering digunakan di dalam kelas adalah di dalam bentuk satu hala iaitu secara pasif. Walaupun kehidupan semula jadi adalah di dalam bentuk 3D, para tenaga pengajar lebih memilih untuk mengajar menggunakan kaedah konvensional 2D kerana cara ini lebih mudah, biasa digunakan, mudah alih dan murah (Kesim, & Ozarslan 2012). Hal ini didapati menjadi kekangan kepada para pelajar untuk memahami dengan jelas jadual berkala kimia. Puncanya kerana mereka tidak dapat melihat kewujudan elemen dan proses yang berlaku seperti pembentukan jirim. Pelajar sukar untuk mengvisualisasikan perkara ini di atas faktor keupayaan imaginasi yang terhad (Cai, Wang & Chiang 2014).

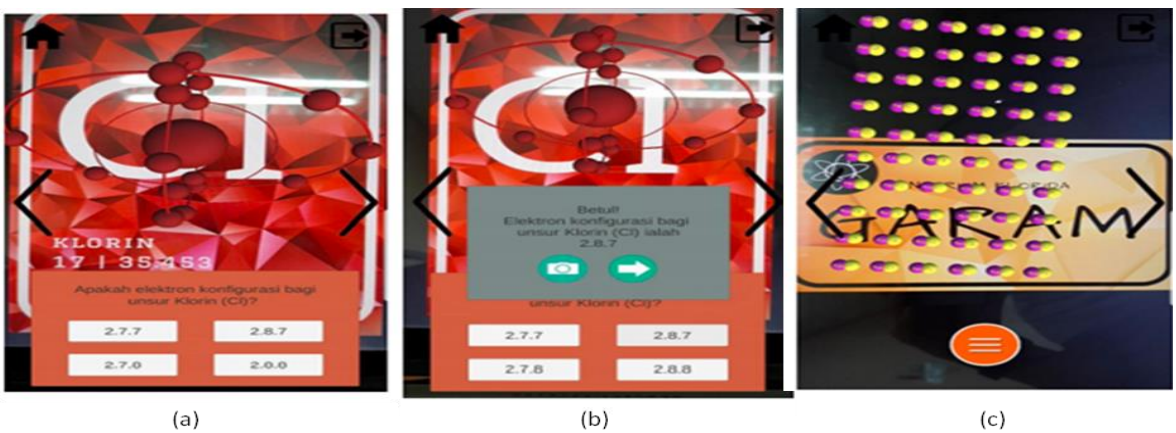
Teknologi AR menjadi pilihan pengajar sebagai satu alternatif tambahan dalam memperkasa pengetahuan pelajar dan kini pembangunan aplikasi berasaskan AR untuk tujuan pembelajaran semakin meluas seperti untuk sains, perbendaharaan kata dan bahasa Arab. Aplikasi mudah alih AR telah dibangunkan untuk kumpulan pertama dalam jadual berkala berasaskan soalan (Abd Majid & Abd Majid), tetapi model 3D untuk atom tidak mengambarkan keadaan sebenar kerana tiada animasi gerakan. Oleh itu, sejajar dengan perkembangan teknologi masa ini, kajian ini fokus pada penggunaan AR dalam sistem pendidikan kerana ianya mampu untuk mengubah corak pembelajaran sedia ada. Ini seterusnya meningkatkan prestasi dan kefahaman pengguna dengan lebih lagi. Antara muka AR menawarkan kelebihan kepada pengguna kerana ianya mampu untuk mengvisualisasikan gambaran bagi tujuan pendidikan (Shelton 2002). Justeru itu, objektif kajian ini adalah mereka bentuk dan membangun sebuah aplikasi mudah alih AR untuk mempelajari jadual berkala kimia berasaskan soalan. Aplikasi ini memudahkan pelajar memahami jadual berkala kimia melalui animasi objek 3D dan video yang seterusnya meningkatkan minat dan penglibatan pelajar dalam proses pembelajaran.

2. Metodologi

Aplikasi mudah alih pembelajaran AR berasaskan soalan dibangunkan menggunakan metodologi ADDIE yang mengandungi lima fasa iaitu analisis, reka bentuk, pembangunan, implementasi dan penilaian. Aplikasi ini dinilai melalui kajian kes di sebuah sekolah menengah di Malaysia. Aplikasi ini direka bentuk untuk menyelesaikan masalah kefahaman pelajar dalam mempelajari unsur dalam jadual berkala. Untuk meningkatkan aspek kognitif dalam pembelajaran, elemen soalan ditambah bagi setiap sesi yang menggunakan teknologi AR. Oleh itu, setiap bahagian utama aplikasi iaitu 1) tindak balas besi dan elemen, 2) struktur atom dan 3) tindak balas elemen dan air, disertai dengan soalan (Rajah 1). Pembangunan aplikasi ini melibatkan perisian Unity 3D, Vuforia, Autodesk 3Ds Max, Audacity dan Adobe Photoshop untuk sistem operasi Android. Rajah 2 (a) menunjukkan contoh struktur atom dalam bentuk model 3D. Model 3D ini mempunyai animasi iaitu petala yang ada pada model atom itu akan berputar mengelilingi nukleus bagi memberikan gambaran sebenar kepada pelajar. Soalan yang berkaitan dengan model 3D tersebut dipaparkan pada skrin yang sama dengan paparan model 3D. Pelajar mendapat maklum balas terus jawapan yang dipilih seperti pada Rajah 2(b). Pelajar juga diberi perkaitan dengan persekitaran seperti garam dengan struktur atomnya (Rajah 2(c)).



Raj. 1. Modul utama dalam aplikasi pembelajaran AR jadual berkala



(a) (b) (c)

Raj. 2. (a) Model 3D atom berserta soalan, (b) Respon untuk jawapan dan (b) Model 3D untuk molekul garam

3. Keputusan dan Perbincangan

Kajian ini menghasilkan aplikasi yang mempunyai beberapa fungsi utama iaitu 1) mewujudkan struktur atom unsur kimia dalam bentuk 3D, 2) menyertakan suara latar yang membekalkan informasi berkaitan elemen 3) menyediakan soalan yang berkaitan dengan topik untuk menguji tahap kefahaman pengguna, 4) menerapkan unsur multimedia seperti teks, audio dan objek 3D dalam reka bentuk grafik yang menarik untuk merangsang minda pengguna, 5) mengolah sebuah modul pembelajaran yang lengkap mengikut piawai yang telah diselaras oleh Kementerian Pelajaran Malaysia bagi topik jadual berkala kimia.

Kajian kes dijalankan di sebuah sekolah menengah yang melibatkan lima orang pelajar. Responden didapati tidak menghadapi masalah dan kelihatan teruja semasa menggunakan aplikasi kerana aliran aplikasi yang jelas, menarik, menyeronokkan dan mencabar minda mereka. Cadangan penambahbaikan bagi aplikasi ini adalah memasukkan unsur daripada kumpulan lain dan mempunyai tahap kesukaran yang berbeza agar dapat menguji

dengan lebih lagi kemampuan pengguna dan kefahaman mereka terhadap topik yang dipelajari. Aplikasi ini juga diharap dapat menyediakan dua kandungan set soalan yang berbeza mengikut sub topik dan juga set soalan yang menyenaraikan secara rawak keseluruhan sub topik berkaitan jadual berkala kimia. Selain model 3D, teks informasi, video yang lengkap yang memaparkan ciri unsur dalam jadual berkala, aplikasi ini juga memaparkan jumlah markah yang diperolehi selepas menjawab soalan yang diberikan. Oleh itu, pelajar dan guru dapat memantau prestasi mereka setelah menggunakan aplikasi ini.

4. Kesimpulan

Aplikasi pembelajaran AR berasaskan soalan telah direka bentuk dan dibangun bagi meningkatkan fokus pelajar pada proses pembelajaran. Ini berikutan pelajar bukan sahaja memerhati model 3D atom yang yang petalanya bergerak, tetapi perlu mengektrak maklumat yang diperlukan berdasarkan pemerhatian bagi menjawab soalan yang diberikan dalam aplikasi. Melalui kajian kes yang dilakukan, pelajar memberi respon positif terhadap aplikasi dari segi minat dan penglibatan. Ini kerana kad imbas yang mewakili setiap unsur dalam jadual berkala yang sebelum ini bersifat statik, ditambah elemen interaktiviti melalui maklumat digital dan soalan. Pemantauan turut dapat dilakukan oleh para guru berdasarkan markah yang diperolehi oleh pelajar setelah selesai sesi pembelajaran menggunakan aplikasi. Ini membolehkan pelajar dan guru mengenal pasti bahagian mana yang sukar difahami kerana aplikasi ini merangkumi tiga sub topik utama dalam mempelajari kumpulan 17. Kajian ini membuktikan yang aplikasi AR yang digabungkan bersama soalan berupaya meningkatkan interaktiviti pelajar terhadap proses pembelajaran.

Penghargaan

Kajian ini disokong oleh geran PDI-2021-025 dan GUP-2020-090.

Rujukan

Abd Majid, N. A. & Abdul Majid, N. 2018. Augmented Reality to Promote Guided Discovery Learning for STEM Learning. Journal on Advanced Science, Engineering and Information Technology, 8 :4-2.
Billinghurst, M., & Duenser, A. 2012. Augmented Reality in the Classroom. Computer, 45 (7), 56-63.
Cai, S., Wang, X. & Chiang, F. K. 2014. A case study of Augmented Reality simulation system application in a chemistry course. Computers in Human Behavior, 37, 31–40. doi:10.1016/j.chb.2014.04.018
Kesim, M. & Ozarslan, Y. 2012. Augmented Reality in Education: Current Technologies and the Potential for Education. Procedia - Social and Behavioral Sciences, 47(222), 297– 302. doi:10.1016/j.sbspro.2012.06.654
Lee, K. 2012. Augmented Reality in Education and Training. Techtrends Tech Trends, 56(2), 13–21. doi:10.1007/s11528-012-0559-3.
Shelton, B. E. 2002. Augmented Reality and Education: Current Projects and the Potential for Classroom Learning. New Horizons for Learning, 9(1): 1–7.

Prediction Model of In-hospital Mortality Post Percutaneous Coronary Intervention (PCI) Using Machine Learning Technique

Rosila Rebo[a,b], Afzan Adam[a*], Azlan Hussin[b]

[a]Center fro Artificial Intelligence Technology, Faculty of Information Science and Technology. National University of Malaysia
[b]Clinical Research Department, National Heart Institute, 145 Jalan Tun Razak, 50400 Kuala lumpur

*Email: afzan@ukm.edu.my

Abstract

The application and development of data science and machine learning plays a vital role in bringing improvisation, innovation and transformation in medicine domain. This empowerment catalyzes the main goal of the study in developing in hospital mortality prediction model for post-Percutaneous Coronary Intervention (PCI) as well as to determine the significant mortality factors. PCI has evolved for four decades in enhancing its effectiveness in treating Coronary Heart Disease (CHD). However post-procedure mortality still haunts the reputation of the modern medical world even at a very low rate. With the growth of PCI National Heart Institute (IJN) data over decade involving 28407 procedures and embedded with robust technology helped achieve this goal. The prediction model has been designed and structured into three tiers of prediction derived from a complete dataset. The advantages of this tiers concept is allows to make an efficient prediction as early as demographic phase and gradually to the intra-procedure phase and post-procedure phase. Thus, it serves as a support medium for clinical decision making at each phase of prognosis. This study began with data exploration and preprocessing that required approximately 80% of effort and time in term to produce high cleaned quality data. This study implemented the filtration Information Gain Ranking for feature selection of the significant factors and sampling SMOTE technique to overcome the problem of extreme imbalance dataset. The dataset was split into training (70%) and testing (30%) and 10-fold cross validation as estimator. Gradient Boosted Decision Tree (GBDT), K Nearest Neighbor (KNN) and Artificial Neural Network (ANN) models were developed. Parameter optimization was implemented at the learning rate for the GBDT and ANN algorithms while the K parameter for the KNN algorithm. Systolic, complications and IABP were the factors with highest information gain for demographic, intra-procedure and post-procedure datasets. Model with SMOTE showed significantly better performance compared to imbalance dataset and no overfitting reported. Overall, GBDT prediction model showed the best performance across demographic, intra-procedure and post-procedure phase then followed by KNN and ANN.

Keywords: PCI prediction; inhospital mortality prediction

1. Introduction

Coronary Heart Disease (CHD) is a non-communicable disease and the most popular type of cardiovascular disease due to its worldwide distribution. CHD contributes the highest death statistic worldwide. Malaysia also ranked 33rd in the world with 29363 (23.1%) deaths in 2014. Critically, Malaysia continues to account a portion of 15.8% from total worldwide of CHD death along with United States, Korea, Japan and others ASEAN country in 2016 (DOSM 2019).

Treatment of CHD via Purcutaneous Coronary Intervention (PCI) has evolved over the past four decades thus giving new hope to CHD patients (Canfield & Totary-Jain 2018). Generally, PCI helps to clear and widen the artery blockage and promote normal blood circulation hence comfort the symptoms dan improve the

patient's quality of life. However, PCI still has the risk of complications like others surgical procedures such as blood clotting, heart attack, bleeding and death after procedure or within 30 days of post procedure.

Therefore, machine learning will be implemented to improve predictive effectiveness, prognosis and improve patient care, as suggested by Gui & Chan (2017). Thus, the main purpose of this study was to develop a PCI mortality prediction model and determine the significant factors of contributing to it. The main challenge however, is the naturally imbalanced data as only 1% death from 2007 to 2016 patients.

Previously, Mohamad & Bee Wah (2019) have developed a post PCI survival model using IJN datasets with the best performance of Naives Bayes algorithm with accuracy was 79.13%, sensitivity was 75.73%, specificity was 82.52%, precision was 81.25% and error rate was 20.87%. Random Under-sampling (RUS) was used as sampling method for imbalance class problem with 300 (1.06%) data used from total 28407 row provided. There are 12 attributes selected based on previous literature review selection and consisted of demography, life style, lipid profile, comorbidities and physical measurements. Therefore, this study aims to make an improvements and use the best approach to develop the best model performance.

1.1. Dataset

Data were taken from IJN with the ethical permission (IJNREC/457/2020), Institutional Review Board and fulfilled the Helsinki Declaration. It consists of 23638 patients with total of 28407 PCI procedures that involving 40244 lesion records since the year 2007 until 2016. Number of attributes were 466 with 44 are demographics data (Dataset A), 126 attributes were from the intra-procedure (Dataset B) and the rest are from post-procedure (Dataset C). Dataset are severely imbalanced as only 1% of death were recorded.

2. Methodology

2.1. Pre-process

Preprocessing phase involving data cleaning that encompasses of elimination of meaningless feature, elimination of feature with more 50% missing value, missing value identification, extraction of useful mining lesion data, data consolidation for table lesion with other table and merged with unique ID, outliers value elimination and missing value and elimination of features with the same meaning. Transformation phase involves the generation of new features, replacing missing data with average values and One-Hot Encoding which gives value of 0 and 1. This phase requires approximately 80% of time and effort.

Table 1 Summary Result for Features Selection for Dataset A, B and C. There are Two Type of Features; Main and One Hot Encoding (OHC) with Total of Feature Selection IG and Matrix Correlation.

Feature Selection	Dataset A		Dataset B		Dataset C	
Type of Feature	Main	OHC	Main	OHC	Main	OHC
Original Number	44	69	126	239	157	286
IG Feature Selection	15	20	53	75	69	96
Main IG Feature Selection	Systolic Group Heart Rate Group MDRD Group		IABP PCI Status STEMI		Complication Cardiogenic Shock IABP	
Total Feature Matrix Correlation	12	13	45	58	60	75
Feature Reduced	72.23%		64.26%		61.78%	

2.2. Feature selection

Features with weightage of information gain above zero were selected. Then the matrix correlation was performed to select only correlation weight that exceeds 0.5. As for the imbalanced data handling, SMOTE techniques (Chawla et. al. 2002) were used. Details on selected features are shown in Table 1.

2.3. Prediction modelling

Dataset treated with SMOTE dataset has achieved better result for all matrix evaluations for both classes compared to the original dataset although there was a slightly decrease in accuracy, specificity, FPR readings, classification error and MSE. Models with SMOTE capabilities predict both survival and death classes with overall accuracy was 98.85%. Thus, this dataset will be used to further improved the classification algorithm.

Two main phase that encompasses training phase and testing phase (Hsieh et al. 2019). Training phase requires 70% while the rest 30% for the testing portion. 10-fold crossvalidation were also arried out as an error estimator for 70% training data thus reducing error variation, estimating accurate performance and to avoid overfitting.

The experiment was conducted on three algorithmic models called Gradient Boosted Decision Tree (GBDT), K-Neighbor Nearest (KNN) and Artificial Neural Network (ANN). Parameter optimization carried out as well to find optimal performance which reduces the loss function for better performance.

3. Result

The best performance of parameter optimization for dataset A, B and C was at learning rate of 0.1 for GBDT**.** GBDT achieved accuracy as 99.30% with 98.87% sensitivity and AUC as 0.998 with misclassification error was 0.7% for dataset A. Detail comparison for each dataset is shown in Fig. 1.

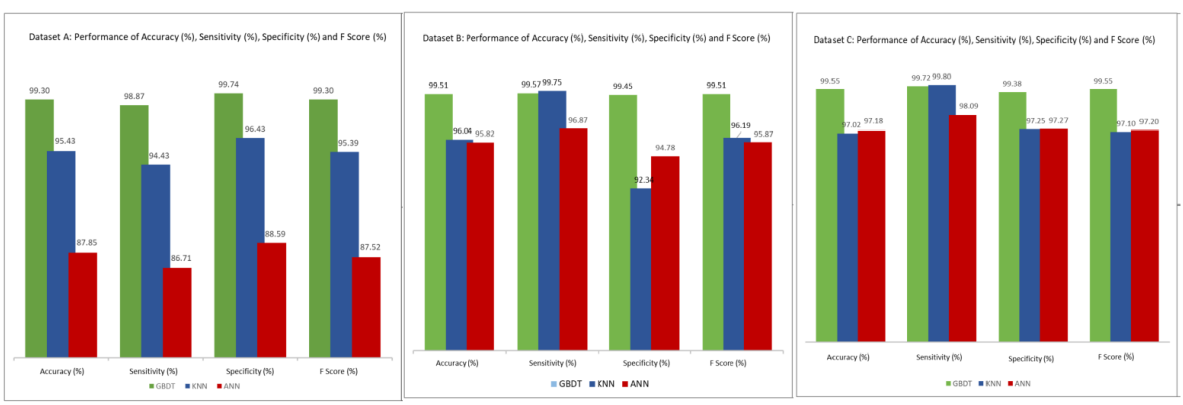


Fig.1.: Comparison of prediction algorithm for dataset A, B and C.

Overall, all algorithms competitively work efficiently above 90.0% for all matrix evaluation. GBDT shows the best performance for all dataset with maximum achievement fall to dataset C although only slightly performance improvement with accuracy was 99.55%, sensitivity was 99.72%, specificity was 99.83%, AUC was 1.000, F Score was 99.55%, FPR was 0.62%, FNR was 0.28%, classification error was 0.45% and MSE was 0.003. Training execution time took 3.5 minutes but still shows fast execution time in line with large features found in dataset C.

4. Discussion

Feature selection using filtration information gain ranking and combination with matrix correlation successfully selects important features for all dataset thus improving model performance. The application of SMOTE techniques improves the performance of the model and be able to predict both survival and mortality classes efficiently. However, as the nature of the data should be very skewed, future research using outliers behaviour detection should be investigated.

Acknowledgements

This research was supported by FTM1 of FTSM. We would like to thank our colleagues from IJN especially PCI Team Clinical Research Department, who provided insight and expertise that greatly assisted the research, as well as physial works while time-off for doing this project.

References

Canfield, J. & Totary-Jain, H. 2018. 40 years of percutaneous coronary intervention: History and future directions. *Journal of Personalized Medicine* 8(4): 1–9. doi:10.3390/jpm8040033

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16(February 2017): 321–357. doi:10.1613/jair.953

Department of Statistics Malaysia. 2019. Perangkaan Sebab Kematian Malaysia 2019. Jabatan Perangkaan Malaysia 57: 1–441

Hsieh, M.-H., Lin, S.-Y., Lin, C.-L., Hsieh, M.-J., Hsu, W.-H., Ju, S.-W., Lin, C.-C., et al. 2019. A fitting machine learning prediction model for short-term mortality following percutaneous catheterization intervention: a nationwide population-based study. *Annals of Translational Medicine* 7(23): 732–732. doi:10.21037/atm.2019.12.21

Gui, C. & Chan, V. 2017. Machine Learning in Medicine 76–78. Retrieved from http://www.uwomj.com/wp--content/uploads/2017/12/vol86no2_28.pdf

Mohamad, R. & Bee Wah, Y. 2019. Prediction of survival status of patients after Percutaneous coronary intervention (PCI) using machine learning techniques. 6(1): 1–46. doi:10.1016/j.surfcoat.2019.125084

Machine Learning in Predicting Cardiovascular Diseases Using ECG Signal

Talal A.A. Abdullah[a], M. Soperi Mohd Zahid[a*], Khaleel Husain[b]

[a]Department of Computer & Information Science, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Malaysia.
[b]Institute of Health and Analytics, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Malaysia.
* Email: msoperi.mzahid@utp.edu.my

Abstract

Statistics have shown that Cardiovascular diseases (CVDs) are the leading cause of death in Malaysia and worldwide. In medicine, evidence-based practice is considered the guiding principle of clinical practice integrating individual clinical expertise and best external evidence in making clinical decisions. However, the demands of diagnoses or generating a prediction for large, heterogeneous populations is nearly impossible for traditional evidence-based methods to keep up with the latest trials and studies in healthcare. Machine learning can be used to discover patterns and associations within massive datasets to assist diagnoses and predict future outcomes. Since the last decades, several methods were reported for automatic ECG beat classifications. In this work, we present a survey of the current state-of-the-art methods used to detect CVDs using ECG signals. It includes the feature selection and machine learning approaches used for automatic detection and decision-making process.

Keywords: Survey; CVD; Healthcare; ECG; Machine Learning.

1. Introduction

Cardiovascular Diseases (CVDs) takes place in the category of fatal diseases resulting in death around the world (Nazlı, Gültepe, & Altural). According to the world health organization, 17.3 million people died from cardiovascular diseases (CVDs) each year, which estimates 31 per cent of all death worldwide (Mendis, Puska, Norrving, Organization, & others, 2011). The rates are remarkably higher in the Middle East, Asia, and Russia than in the rest of the world (Alizadehsani, Abdar, et al., 2019; Zipes, Libby, Bonow, Mann, & Tomaselli, 2018). In CVDs, angiography is widely used in cardiology by clinicians as it considered the most precise method (Kim & Choi, 2015; Kim et al., 2015; Tsipouras et al., 2008). It is, however, an invasive and costly procedure, and it may lead to various complications (Alizadehsani et al., 2018; Mnih et al., 2015). Electrocardiogram (ECG/EKG) is another most common diagnostic tool for recording physiological heart activities during a specific period. It is estimated that over 300 million ECGs are recorded worldwide per year (Holst, Ohlsson, Peterson, & Edenbrandt, 1999), and the number keeps growing. ECG is a non-invasive and uncostly tool that can aid in diagnosing many cardiovascular abnormalities, such as atrial fibrillation (AF), premature contractions of the atria (PAC) or ventricles (PVC), congestive heart failure (CHF), and myocardial infarction (MI) (Hong, Zhou, Shang, Xiao, & Sun, 2020). However, ECG signals are noisy sensitive and need a human expert to understand what they mean and are most likely to be misunderstood. Therefore, computeraided interpretation of ECGs has become more crucial, especially in low-income and middle-income countries where experienced cardiologists are scarce (Organization & others, 2014). Although expert features can automatically be extracted using some computer-based programs, they are still insufficient. The main reason is that they are limited by human expert knowledge and data quality (Guglin & Thatai, 2006; Schläpfer & Wellens, 2017; Shah & Rubin, 2007). Therefore, researchers are working on alternative methods to extract ECG signal features that do not require an explicit feature extraction by human experts such as machine learning and deep learning. Machine Learning (ML) is tremendously used recently in many areas, such as speech recognition and image processing. The revolution in industrial technology proves the great success of machine learning and its applications in analyzing

intricate patterns, which presented in a wide range of applications in various fields, including healthcare (Adadi & Berrada, 2018). The rest of this survey as structured as followed. In Section 2 we introduce the ECG structure and components, and the diseases that can aid to diagnose. Section 3 describes the related state-of-art works used to detect CVDs using ECG signals. The widely used MACHINE LEARNING algorithms, feature selections, and datasets are discussed in this section.

2. Electrocardiogram Signal (ECGs)

Electrocardiogram Signal or ECGs is a non-invasive, easy to acquire, and uncostly diagnose tool used for recording the heart's physiological activities of the heart (Hong et al., 2020; Sahoo, Dash, Behera, & Sabut, 2020). ECGs can aid in diagnosing many cardiovascular abnormalities such as Premature contractions of the atria (PAC) or ventricles (PVC), Atrial fibrillation (AF), Myocardial infarction (MI), and Congestive heart failure (CHF) (Holst et al., 1999). The standard ECG has 12 leads; six of them are placed on the arms and legs of the patient that are labelled as "Limb Leads", and the other six leads are placed on the torso (precordium), and they labelled as "Precordial Leads". The limb leads are called lead I, II, III, aVL, aVR and aVF and the precordial leads are called leads V1, V2, V3, V4, V5 and V6. A normal ECG contains waves, intervals, segments and QRS complex, (Fig. 1) (Sahoo et al., 2020).

Waves represent a positive or negative deflection from the baseline that indicates a specific electrical event. The waves on an ECG include the P wave, Q wave, R wave, S wave, T wave and U wave. Intervals represent the time between two specific ECG events. The intervals commonly measured on an ECG include the PR interval, QRS interval (also called QRS duration), QT interval and RR interval. Segment is the length between two specific points on an ECG supposed to be at the baseline amplitude (not negative or positive). The segments on an ECG include the PR segment, ST-segment and TP segment. Complex represents the combination of multiple waves grouped together. The only main complex on an ECG is the QRS complex.

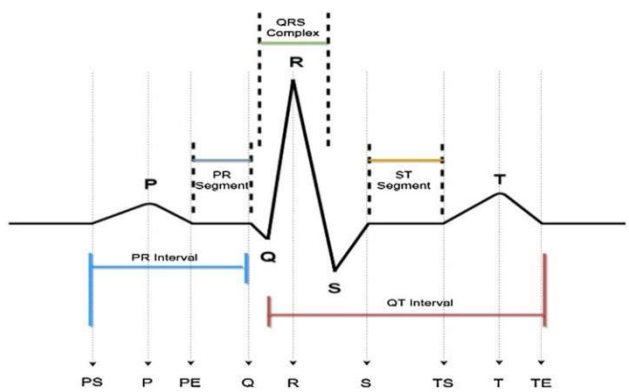


Fig. 1 Electrocardiogram Sample (ECG)

The main part of an ECG contains a P wave, QRS complex and T wave. The P wave indicates atrial depolarization. The QRS complex consists of a Q wave, R wave and S wave and represents ventricular depolarization. Finally, the T wave comes after the QRS complex and indicates ventricular repolarization.

3. Related Works

In (Kiranyaz, Ince, & Gabbouj, 2015), the author proposed a real-time ECG classification and monitoring system. A 1D-CNN algorithm trained on MIT-BIH dataset to diagnose CVDs. The model achieved 98.9% accuracy, 95.9% sensitivity, and 99.4% specificity. The MIT-BIH dataset was used in (Zubair, Kim, & Yoon, 2016) to train a CNN algorithm to diagnose CVDs and automatically classified ECG beats into five different normal beat classes Supraventricular ectopic beat, Ventricular ectopic beat, Fusion beat, Unknown beat. The proposed model achieved 92.7% overall accuracy.

Two- and five-seconds durations of ECGS signal segments from Fantasia and St.Petersburg datasets are used to train a CNN algorithm to predict CAD in (Acharya et al., 2017). The CNN model structures comprising of four convolutional layers, four max-pooling layers and three fully connected layers. The proposed model is capable of differentiating between normal and abnormal ECG with 94.95% precision, 93.72% sensitivity and 95.18% specificity for Net 1 (two seconds) and 95.11% accuracy, 91.13% sensitivity and 95.88% specificity for Net 2 (5 s). Different kernels of Least Squares-Support Vector Machine (LS-SVM) are used in (Kumar, Pachori,

& Acharya, 2017) on ECG signals of 40 normal people and 7 CAD patients. Student's t-test method and Kruskal–Wallis statistical test are applied to check the extracted features' discrimination ability. The developed model achieved 0.99% for accuracy, sensitivity, and specificity, respectively.

Table 1. Related Works in Predicting CVDs using ECGs

Approach	Dataset	Method	Feature Selection	Output	Performance%
Kiranyaz et al. (2015)	MIT-BIH	CNN	NA	Classifying ECG	AC = 0.98, SEN = 0.95, SPC = 0.99
Zubair et al. (2016)	MIT-BIH	CNN	NA	Classifying ECG	Ac = 0.93
Acharya et al. (2017)	Fantasia, St.-Petersburg	CNN	NA	Diagnosing CAD	AC = 94.95, SEN = 93.72, SEN = 95.18
Kumar et al. (2017)	Fantasia, St.-Petersburg	LS-SVM	NA	Diagnosing CAD	AC = 0.99, SEN = 0.99, SPC = 0.99
Tan et al. (2018)	PhysioNet	LSTM, CNN	NA	Diagnosing CAD	AC = 99.85
(Andersen et al., 2019)	MIT-BIH	CNN, RNN	NA	Classifying ECG	SEN = 0.98, SPC =0.96
Sharma and Acharya (2019)	PhysioNet, St.Petersburg	GSVM	OTFC, BWFB	Identifying CAD	AC = 99.53, SEN = 98.64, SPC = 99.70
(Butun et al., 2020)	Fantasia, St.-Petersburg	CapsNet	NA	Detecting CVD	2sAC=99.44, 5sAC = 98.62
(Al-Zaiti et al., 2020)	EMPIRE study	LR, GBM, ANN	NA	Predicting CAD syndrome	SEN =77.0, SPC = 76.0

Abbreviations: **CNN**, Conventional Neural Networks; **LS-SVM**, Least-Squares Support-Vector Machine; **LSTM**, Long Short-Term Memory; **RNN**, Recurrent Neural Network; **GSVM**, Gaussian Support Vector Machine; **CapNet**, Capsule Neural Network; **LR**, Logistic Regression; **GBM**, Gradient Boosting Machine; **ANN**, Artificial Neural Network; **NA**, Not Available; **OTFC**, Optimally Time-Frequency Concentrated; **BWFB**, Biorthogonal Wavelet Filter Bank; **CAD**, Coronary Artery Disease ;**AC**, Accuracy; **SEN**, Sensitivity; **SPC**, Specificity.

In (Tan et al., 2018), authors implement a long short-term memory (LSTM) network with (CNN) to automatically diagnose CAD ECG signals. Using Fantasia dataset for normal data and ST-Petersburg for CAD patients, the model achieved 99.53%, 98.64%, 99.70% accuracy, sensitivity, and specificity, respectively. A real-time approach for automatic detection of atrial fibrillation (AF) in long-term electrocardiogram (ECG) is developed in (Andersen, Peimankar, & Puthusserypady, 2019). A combination of CNN and RNN algorithms trained using MIT-BIH dataset and achieved 98.0% sensitivity and 96.0% specificity. For automatically identifying CAD, (Sharma & Acharya, 2019) proposed the use of optimally time-frequency concentrated (OTFC) even-length biorthogonal wavelet filter bank (BWFB). The model was trained on a 10-fold cross-validation technique and Gaussian Support Vector Machine (GSVM) algorithm to diagnose CAD. The average sensitivity and specificity obtained are 0.98% and 0.99%, respectively, with the Matthews correlation coefficient(MCC) of 98.0%. 1D-CADCapsNet (Butun, Yildirim, Talo, Tan, & Acharya, 2020) provides an automated detection for CAD from ECG signals using Capsule Network algorithm (CapsNet). The proposed model was trained on two seconds (95,300) and five second-long (38,120) ECG segments from Fantasia and ST-Petersburg data sets. The 1D-CADCapsNet model yielded a 5-fold diagnosis accuracy of 99.0% and 98.0% for two and five-second ECG signal groups, respectively. (Al-Zaiti et al., 2020) developed machine learning-based methods for predicting underlying acute myocardial ischemia in patients with chest pain. The model trained and tested multiple classifiers on two independent prospective patient cohorts using 554 temporal-spatial features of the 12-lead ECG.

3.1. Machine Learning

According to the literature, SVM and CNN are widely used in CVDs diagnosis using ECGs data. Both SVM and CNN are considered powerful algorithms, and they achieved very high accuracy rates compared with other methods (Ahmad et al., 2014). The main reason for SVM significant performance is the feature dimensionality independency which makes it immune from the "curse of dimensionality". SVM works well with a clear margin of separation between classes. The hyperplane is affected by only the support vectors. Thus, outliers have less impact, which makes it particularly efficient for classifying complex but small or medium-sized datasets (Li, Bhanu, & Krawiec).

Furthermore, different SVM classifiers can be constructed using different kernels (polynomial, RBF, linear) to solve leaner or non-leaner problems. However, selecting an appropriate kernel for a. given problem is still under research. Besides, SVM does not work perfectly in a noisy or large dataset, and it takes a long time in training the model (Bisong, 2019). In contrast, CNN can work perfectly in a noisy and large dataset because of its capability of finding optimal temporal features better than other machine learning algorithms.

CNN can learn complex features by preserving the spatial relationship between the feature (Bisong, 2019). Reducing the number of weights needed for training the network is another significant advantage of CNN. However, both SVM and CNN are High computational costs and time and memory consuming (Alizadehsani, Roshanzamir, et al., 2019).

A small number of databases are available in the public domain, which explicitly studied in most previous works. Further, most of these public databases have a small sample size and unbalanced data. MIT-BIH, ZAlizadeh Sani, Cleveland, and Hungarian datasets are the most popular public dataset. The main reason traditional machine learning algorithms are significantly well performed is that most of the sample size of these public domain datasets, accept MIT-BIH, is between 303 and below with few features. As seen from the literature, a feature selection method has been used in most structured data works, and it has a significant impact on model performance.

In signal data, a combination of the Fantasia dataset and St.-Petersburg dataset are widely used. The ECG signals were retrieved from Fantasia (for Normal) and St.-Petersburg Institute (for CAD). However, most datasets in this field are either very small in size or inaccessible to the public. One reason traditional ML has worked sufficiently well in previous years in using ECGs to diagnose cardiology is that experts are carefully designed feature extraction methods. Also, handcrafted features such as statistical measures from the ECG beats and the RR interval positively influence traditional ML's performance. Recently, using DL to diagnose CVDs from ECG signals has gained much interest due to its simplicity and reduced dimensionality compared to imaging data (Strodthoff, Wagner, Schaeffter, & Samek, 2020).

4. Summary

Signal data and especially ECG signals have gained a lot of interest in the research community with respect to the other data type. ECG is a widely available and great tool used to diagnose CVDs with high efficiency from a medical aspect. However, it is noisy sensitive and can be misleadingly interpreted by doctors. ML and DL incredibly proved their ability to insight hiding information and provide simplicity in handling such problems. CNN algorithms particularly have achieved high efficiency and outstanding performance in diagnosing CVDs using ECG signals. This survey introduced the ECG structure and components, and the diseases that can aid to diagnose. The related state-of-art works used to detect CVDs using ECG signals is discussed as well as the widely used ML algorithms, feature selections, and datasets.

Acknowledgment

The authors acknowledge the support of this research by the Yayasan Universiti Teknologi PETRONAS Fundamental Research Grant (YUTP-FRG) under Grant 015LC0-244

References

Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., & Chua, C. K. (2017). Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. *Knowledge-Based Systems, 132*, 62–71-62–71.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160-52138–52160.

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews, 33*, 102–109-102–109.

Al-Zaiti, S., Besomi, L., Bouzid, Z., Faramand, Z., Frisch, S., Martin-Gill, C., . . . Sejdić, E. (2020). Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature communications, 11*(1), 1–10-11–10.

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., . . . Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine, 111*, 103346-103346.

Alizadehsani, R., Hosseini, M. J., Khosravi, A., Khozeimeh, F., Roshanzamir, M., Sarrafzadegan, N., & Nahavandi, S. (2018). Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. *Computer Methods and Programs in Biomedicine, 162*, 119–127-119–127.

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., . . . Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific data, 6*(1), 1–13-11–13.

Andersen, R. S., Peimankar, A., & Puthusserypady, S. (2019). A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications, 115*, 465–473-465–473.

Bisong, E. (2019). *Building Machine Learning and Deep Learning Models on Google Cloud Platform*: Springer.

Butun, E., Yildirim, O., Talo, M., Tan, R.-S., & Acharya, U. R. (2020). 1D-CADCapsNet: One dimensional deep capsule networks for coronary artery disease detection using ECG signals. *Physica Medica, 70*, 39–48-39–48.

Guglin, M. E., & Thatai, D. (2006). Common errors in computer electrocardiogram interpretation. *International journal of cardiology, 106*(2), 232–237-232–237.

Holst, H., Ohlsson, M., Peterson, C., & Edenbrandt, L. (1999). A confident decision support system for interpreting electrocardiograms. *Clinical Physiology, 19*(5), 410–418-410–418.

Hong, S., Zhou, Y., Shang, J., Xiao, C., & Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 103801-103801.

Kim, H. Y., & Choi, J.-H. (2015). TCTAP A-084 Lesion-Specific Myocardial Mass: A New Index for Diagnosis and Treatment of Coronary Artery Disease. *Journal of the American College of Cardiology, 65*(17 Supplement), S43–S44-S43–S44.

Kim, H. Y., Kim, E. k., Kim, S. M., Song, Y. B., Hahn, J.-Y., Choi, S.-H., . . . others. (2015). Fractional myocardial mass: a new index for diagnosis and treatment of coronary artery disease. *Journal of the American College of Cardiology, 65*(10S), A1269–A1269-A1269–A1269.

Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering, 63*(3), 664–675-664– 675.

Kumar, M., Pachori, R. B., & Acharya, U. R. (2017). Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals. *Biomedical signal processing and control, 31*, 301–308-301–308.

Li, R., Bhanu, B., & Krawiec, K. (2007). *Hybrid coevolutionary algorithms vs. SVM algorithms.* Paper presented at the Proceedings of the 9th annual conference on Genetic and evolutionary computation.

Mendis, S., Puska, P., Norrving, B., Organization, W. H., & others. (2011). *Global atlas on cardiovascular disease prevention and control*: World Health Organization.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . others. (2015). Humanlevel control through deep reinforcement learning. *nature, 518*(7540), 529–533-529–533.
Nazlı, B., Gültepe, Y., & Altural, H. Classification of Coronary Artery Disease Using Different Machine Learning Algorithms.
Organization, W. H., & others. (2014). *Global status report on noncommunicable diseases 2014*: World Health Organization.
Sahoo, S., Dash, M., Behera, S., & Sabut, S. J. I. (2020). Machine learning approach to detect cardiac arrhythmias in ecg signals: a survey.
Schläpfer, J., & Wellens, H. J. (2017). Computer-interpreted electrocardiograms: benefits and limitations. *Journal of the American College of Cardiology, 70*(9), 1183–1192-1183–1192.
Shah, A. P., & Rubin, S. A. (2007). Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of electrocardiology, 40*(5), 385–390-385–390.
Sharma, M., & Acharya, U. R. (2019). A new method to identify coronary artery disease with ECG signals and time-Frequency concentrated antisymmetric biorthogonal wavelet filter bank. *Pattern Recognition Letters, 125*, 235–240-235–240.
Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2020). Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *arXiv preprint arXiv:2004.13701*.
Tan, J. H., Hagiwara, Y., Pang, W., Lim, I., Oh, S. L., Adam, M., . . . Acharya, U. R. (2018). Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Computers in Biology and Medicine, 94*, 19–26-19–26.
Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., & Michalis, L. K. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine, 12*(4), 447–458-447–458.
Zipes, D. P., Libby, P., Bonow, R. O., Mann, D. L., & Tomaselli, G. F. (2018). *Braunwald's Heart Disease EBook: A Textbook of Cardiovascular Medicine*: Elsevier Health Sciences.
Zubair, M., Kim, J., & Yoon, C. (2016 2016). *An automated ECG beat classification system using convolutional neural networks.* Paper presented at the 2016 6th international conference on IT convergence and security (ICITCS).

Image Compression in Digital Pathology

Goh Jee Yuan[a*], Afzan Adam[b], Zaid Alyasseri[c]

[a,b,c] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, UKM 43600 Bangi, Selangor, Malaysia

* Email:P102192@siswa.ukm.edu.my

Abstract

The rise of digital pathology has brought about great conv enience to both the public and medical practitioners. The process of diagnosing a patient has now become much more efficient with the aid of digital pathology, which enable the managing and diagnosing of pathology slides digitally. At the same time, this has also led to a surge in the amount and size of image data generated, as pathology slide images are high in density and zoomable. As the amount of image data increases exponentially, it leads to the problem of insufficient storage space and low transfer efficiency across devices, due to the image complexity and large image size. This research aims to review the latest image compression algorithm, in an attempt to reduce the size of pathology slide images. These microscopic images can get very large as the images can have very large dimensions or requires very fine attention to small details. The objective of this research is to investigate the performance of some common image compression algorithms when tested on high-density pathological image datasets.

Keywords: Digital Pathology; Image Compression; High-density Images

1. Introduction

The digitization of images has brought upon great convenience to our daily lives as images can now be digitized and stored in devices as files that can be easily and readily accessed. These digital images can also be easily shared to other devices over the internet to remote places in a matter of minutes. But with the digitization process comes the need for storage. All files, whether it be images, videos, texts, etc. takes up space to store, in the form of data bytes (Buchholz 1962). As digital imaging advances in quality, the resolution, colour depth and detail of such images greatly improves (Tyagi 2018). As such, more bytes are required to preserve the data of each individual image, which causes an exponential increase in image size. This results in the problem of insufficient storage on devices and slow transfer speeds across devices, especially for large images due to the large volume of data to be transferred (Wu 2006).

The aim of this research is to look into image compression as a means to reduce image size for ease of storage and transfer, while at the same time maintaining a similar visual perception in the resulting image when compared to the original image. Image compression is a form of data compression that is applied on digital images to reduce the size of images for storage and transfer. Compression is achieved through the application of an algorithm on the digital image data that modifies the data in a way that the data retains a similar visual perception but is reorganized in a more storage cost effective solution hence reducing the size of the digital image (Rahman 2019).

2. Methodology

2.1. Data acquisition

The focus of this research is to investigate the performance of some common image compression algorithms when tested on high-density pathological image dataset. The dataset is open source and obtained ethically under

America's National Cancer Institute's (NCI) Genomic Data Commons (GDC) portal. A total of 85 pathological images were randomly selected from the GDC portal to form the dataset for this research.

2.2. Algorithm testing

For this research, four algorithms have been selected for testing. The four algorithms selected for this research are:

- Discrete Cosine Transform (DCT), a lossy image compression algorithm (Maru 2020)
- Discrete Wavelet Transform (DWT), a lossy image compression algorithm (Nain 2020)
- Huffman Encoding, a lossless image compression algorithm (Kai-Meng 2020)
- Set Partitioning in Hierarchical Trees (SPIHT), a near-lossless image compression algorithm (Fangfang 2020)

These four algorithms were selected for this research as the first three algorithms represent algorithms that being widely used in image compression of other types of image (Mantoro 2017). By comparing the SPIHT algorithm against these algorithms, it would be a fair head-to-head comparison and representation of how these algorithms would perform in a daily use case (Rahman 2019).

2.3. Performance Evaluation

There can be numerous ways to evaluate a compression algorithm, whether it be through compression speed, a measure of how fast an algorithm processes an image and compresses it, through compression ratio, a measure of how much more compact the compressed image is when compared to the original file in size, through image quality, a measure of how similar the reconstructed image is when compared to the original image (Hussain 2018). These measures serve as a quantifiable measurement to evaluate the efficiency of an image compression algorithm.

In order to measure the compression speed of an algorithm, encoding time is used, which refers to the amount of time taken, in seconds, for the encoder of a compression algorithm to encode and compress the data in the original input image into the outputted compressed data (Sharma 2017).

Compression Ratio (CR) refers to the ratio of the number of bits between the original uncompressed file and the compressed file. It is used to measure how much space has been compressed by the algorithm in the image. For example, a CR of 4:1 signifies that the compressed image is only about 1/4th of the size of the original image (Rahman 2019). The equation to determine CR is as illustrated as follows:

$$Compression\ Ratio\ (CR) = \frac{Number\ of\ Bits\ in\ Original\ Image}{Number\ of\ Bits\ in\ Compressed\ Image} \quad (1)$$

Although both are used to quantify the amount of compression that is applied to a file, CR is not to be confused with compression rate, measured in Bits Per Pixel (BPP), which represents the number of bits in average, that is needed to represent a pixel in an image (Hussain 2018). The equation to measure BPP is as illustrated as follows:

$$Bits\ Per\ Pixel\ (BPP) = \frac{Number\ of\ Bits\ in\ Image}{Number\ of\ Pixels\ in\ Image} \quad (2)$$

There are multiple ways to measure image quality when attempting to evaluate an image compression algorithm. Commonly used methods to measure image quality would be Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR is measured in decibels (dB) and represents the quality of an image relative to the size of the error, where a high PSNR value represents a low measure of error in the reconstructed image when compared to the original (Hussain 2018). The equation to measure PSNR is as illustrated as follows, where n represents the number of bits that represent the pixel:

$$PSNR = 10\frac{((2^n)-1)^2}{MSE} \quad (3)$$

Where MSE represents the Mean Squared Error, which is the average value of the combined square of errors, or differences in each pixel between the original and compressed image (Sharma 2017). The equation to measure MSE is as illustrated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (4)$$

However, there are certain instances in which the MSE value of an image becomes inconsistent and hence a secondary measure of image quality, SSIM is introduced. SSIM measures image quality degradation between the original and compressed image by observing the perceivable structural differences between the two images and serves as an alternative besides PSNR for image quality measurement (Wang 2004). The equation to measure SSIM is as illustrated as follows with μ_x, μ_y representing the average of x and y, and σ_x, σ_y representing the variance of x and y, and σ_{xy} representing the covariance of x and y:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \quad (5)$$

3. Findings

After compression has been applied on the dataset, the Encoding Time, CR BPP , PSNR and SSIM) for each image using each algorithm was recorded in a table. The average value for these parameters were then calculated and tabulated in the table presented below.

Table 1. Average compression results for each algorithm on a sample of 85 images

Algorithm	Time (s)	CR	BPP	PSNR (dB)	SSIM (%)
Discrete Cosine Transform (DCT)	2.1845	57.0007	0.1420	37.2448	95.97
Discrete Wavelet Transform (DWT)	1.2088	25.2386	0.3500	9.0928	55.86
Huffman Encoding	1.4342	1660.8528	0.0145	30.0957	91.75
Set Partitioning in Hierarchical Trees (SPIHT)	14.2125	480.8744	0.0523	35.1604	95.78

From the table above, we can observe that the DWT algorithm produces the lowest performance out of all four algorithms that were tested. While the other three each excels in certain criteria. DCT produces a fast and a good quality compression, as seen with high PSNR and SSIM numbers, but with a lower CR compared to Huffman and SPIHT. Huffman on the other hand, is also fast and produces a much more compact compression as seen in the high CR numbers but produces a lower quality compression when compared to DCT and SPIHT. Amongst all the algorithms that were tested, SPIHT produced a very good overall result, being quite similar in performance with the top performing algorithm in each aspect, resulting in a high CR, PSNR and SSIM compression performance, with only one caveat, and that is the amount of time taken due to the complexity of the algorithm resulting in slow encoding speeds.

4. Conclusion

From the data that was gathered through this research, it can be clearly seen that image compression serves as an effective way to reduce the size of images, whether it be for more efficient storage, or for increased transfer efficiency. Based on the results, it can be said that the SPIHT algorithm performs well in all aspects besides encoding time. This can be a point to focus on in coming research to optimize the efficiency of the SPIHT

algorithm or to explore other options in attempts to shorten down the encoding time of the algorithm while maintaining its current performance.

Acknowledgements

This research was supported by KPT grant [FRGS/1/2019/ICT02/UKM//02/6]. I would also like to express my gratitude to Ts. Dr. Abdul Hadi Abd Rahman and Assoc. Prof. Dr. Azizi Abdullah for providing me with much support, from technical knowledge to advice and tips.

References

A.J. Hussain, Ali Al-Fayadh, & Naeem Radi. (2018). Image Compression Techniques: A Survey in Lossless and Lossy algorithms. Neurocomputing.

David Wu, Damian M. Tan, Marilyn Baird, John DeCampo, Chris White & Hong Ren Wu. (2006). Perceptually Lossless

Medical Image Coding. IEEE Transactions on Medical Imaging, 25(3), 335-344.

Fangfang Li, Sergey Krivenko, Vladimir Lukin. (2020). A Fast Method for Visual Quality Prediction and Providing in

Image Lossy Compression by SPIHT. Integrated Computer Technologies in Mechanical Engineering – 2020, 17-29.

Garima Nain, Ashish Gupta, Rekha Gupta. (2020). DWT Based Compression Algorithm on Acne Face Images. International Conference on Intelligent Computing and Smart Communication 2019. 985-993.

Kai-Meng Chen, Chin-Chen Chang. (2021). High-capacity separable reversible data-Hiding method in encrypted images

based on block-level encryption and Huffman compression coding. Connection Science.

Md. Atiqur Rahman, Mohamed Hamada. (2019). Lossless Image Compression Techniques:A State-of-the-Art Survey.

Symmetry, 11(1274), 1-22.

Rashmi Sharma, & Priyanka. (2017). A Review on Study and Analysis of Various Compression Techniques. International

Journal of Innovative Science and Research Technology, 2(2), 4-9.

Teddy Mantoro, Fifit Alfiah. (2017). Comparison Methods of DCT, DWT and FFT Techniques Approach on Lossy Image

Compression. 2017 International Conference on Computing, Engineering, and Design (ICCED).

Umesh Maru, Gajendra Sujediya, Yashika Saini. (2020). Color Image Encryption and Compression Using DCT in Joint

Process. Proceedings of International Conference on Communication and Computational Technologies, 97-111.

Vipin Tyagi. (2018). Understanding Digital Image Processing. Florida: CRC Press

Werner Buchholz. (1962). Planning A Computer System. London: McGraw-Hill Book Company Inc.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, & Eero P. Simoncelli. (2004). Image Quality Assessment: From Error

Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13(4), 1-14.

Impact of Bidirectional LSTM Layer Variation on Cardiac Arrhythmia Detection Performance

Shahab Ul Hassan [a], Mohd Soperi Mohd Zahid [a*], Khaleel Husain [b]

[a]*Department of Computer & Information Sciences, Universiti Teknologi PETRONAS,* Seri Iskandar, Malaysia
[b]*Institute of Health and Analytics, Universiti Teknologi PETRONAS,* Seri Iskandar, Malaysia
* Email: msoperi.mzahid@utp.edu.my

Abstract

Cardiac arrhythmia is responsible for significant fatalities and is one of the critical cardiovascular diseases. Prediction of arrhythmia at the right time can save lives and is of considerable importance to current researchers and clinicians. Deep learning models are recently being used to analyze the electrocardiogram (ECG) signal and arrhythmia prediction. Bidirectional long short-term memory (BiLSTM) algorithms are commonly used deep learning techniques for arrhythmia classification. However, there is a lack of study that analyses the impact of BiLSTM layer variation on the performance of the deep learning model for arrhythmia prediction. In this paper, the performance of BiLSTM algorithms for arrhythmia classification with varying numbers of BiLSTM layers is analyzed. Specifically, the MIT-BIH arrhythmia dataset is used, and the performance is measured in terms of accuracy, recall, precision, and specificity. Evaluating the performance of these algorithms will aid in the creation of a useful BiLSTM model for arrhythmia prediction that uses the optimum number of BiLSTM layers.

Keywords: Arrhythmia; BiLSTM; ECG; Accuracy; Specificity; Precision; Recall.

1. Introduction

Cardiovascular diseases (CVDs) are a major threat to human health because of their high morbidity and mortality rates. The electrocardiogram (ECG) is a safe and effective, transthoracic diagnostic technique that offers a wealth of details on diagnosing and treating cardiovascular diseases (Mukhopadhyay et al., 2012). It is commonly used in tracking the operation of heartbeats. Arrhythmias normally occur in a particular situation and are represented in the ECG by an irregular, sluggish, or a sign of a quick erratic heartbeat. The detection and ECG arrhythmias signal classification may provide valuable medical diagnosis information based on the different forms of arrhythmias related patterns (Lay-Ekuakille et al., 2013). Manually analyzing ECG features is time-consuming and repetitive, and hence it is important to build an automated ECG analysis algorithm. Several cardiac arrhythmia prediction algorithms have been proposed in the last few years.

Recently, ECG signal analysis has been successfully applied using deep learning-based approaches. Deep learning architectures are usually made up of restricted Boltzmann machines, Deep belief networks (DBNs) and deep Boltzmann machines (DBMs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). The long short-term memory (LSTM) network is a type of RNN widely used to analyze time-series data. It can efficiently preserve historical information and acquire long-term text dependency information. ECG arrhythmia has been detected in the past using LSTM (Lay-Ekuakille et al., 2013, Hassan et al., 2020, Husain et al., 2021, Oh et al., 2018, Hou et al., 2019). The bidirectional long short-term memory (BiLSTM) algorithm is an extension of the LSTM network that uses a two-way learning approach. In this paper, the performance of the BiLSTM-based model for ECG arrhythmia classification for the varying number of Bi-

LSTM layers is analyzed.

2. Background

In most artificial intelligence applications involving sequential data or one-dimensional signals, RNN architecture is widely used for prediction, detection, and recognition. This neural network is capable of effectively retaining historical data and learning long-term text dependence information. The LSTM is a variation of the RNN architecture that would handle input and output of different lengths. Because of this asset, it has recently shown potential in sequence learning applications (Lay-Ekuakille et al., 2013)..

2.1 Bidirectional LSTM:

BiLSTM is a network extension of the LSTM. Figure 1 shows the general architecture of BiLSTM, which entails replicating the network's first layer; the input sequence is then fed to the first layer as is, with a reversed copy going to the replicated layer. A BiLSTM can be trained using all possible input data from the past and future of a time step. The forward states (positive time direction) are controlled by a portion of the state neurons, while the backward states are controlled by the other half (negative time direction). The BiLSTM layer's output updates a collection of global features heavily affected by nearby inputs and localized (Cui et al., 2018).

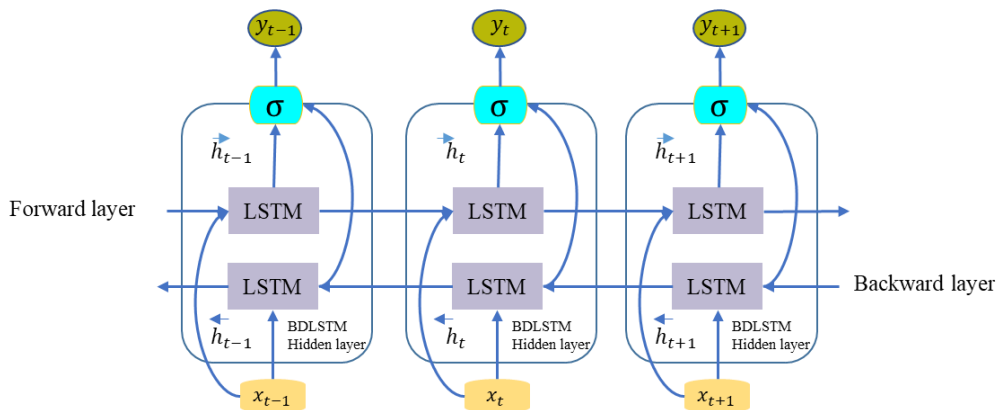


Fig. 1. General Architecture of Bidirectional LSTM Network Framework.

2.2 Related Work:

There has been considerable work on LSTM models for arrhythmia classification. Work by Sharma et al., 2020 utilizes deep LSTM networks to detect and identify four different types of abnormalities in ECG signals. Another model was proposed in Yildirim (2018) for deep BiLSTM network-based wavelet sequences. Here, for the generation of ECG signal sequences, a wavelet-based layer is utilized. Furthermore, the performance of unidirectional LSTM and BiLSTM algorithms was compared. An automatic arrhythmia classifier was developed in He et al., 2019 by combining residual convolutional neural networks (CNN) and BiLSTM layers for the feature extraction from raw ECG signals. The feature vector that is obtained after feature merging is trained for final classification. Finally, work in Xu et al., 2020 proposed a hybrid CNN-BiLSTM architecture for ECG heart signal classification for diagnostic purposes. Specifically, two CNN layers, two BiLSTM layers, and two fully connected layers were utilized in the proposed architecture.

3. Methodology

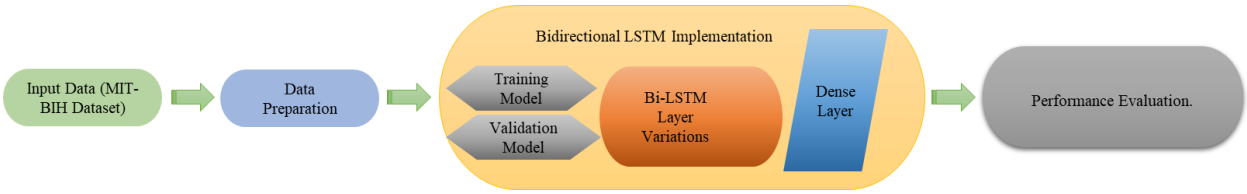


Fig. 2. Flowchart of proposed research.

3.1 MIT-BIH Arrhythmia Dataset:

We used the well-known arrhythmia database from the Massachusetts Institute of Technology (MIT). The MIT-BIH Arrhythmia Dataset is the original dataset and publicly available (Moody and Mark, 2005). The MIT-BIH arrhythmia contains a total of 48 records, each lasting approximately 30 minutes, and obtained between 1975 and 1979 from a two-channel ambulatory device (Moody et al., 2001).

3.2 Data Preparation:

The method of cleaning and converting raw data prior to processing and analysis is known as data preparation. Before processing, it is a critical stage that always includes reformatting data, making data corrections, and merging data sets to refine data. Data preparation ensures data integrity, resulting in accurate observations (Picon et al., 2019).

3.3 BiLSTM Implementation:

BiLSTM layers are then implemented. Specifically, five cases are considered where the number of BiLSTM layers is varied from 1 to 5. In addition, out of 48 records, data of 36 patient's record was utilized by training module, and 12 patient's record was used for validation. Also, a random approach was utilized to split the data into training and validation modules.

3.4 Performance Evaluation:

The performance of the proposed BiLSTM model was measured in terms of accuracy, recall, precision, and specificity. The four metrics are defined as follows:

- **Accuracy** can be characterized as "the degree to which a measurement's outcome conforms to the right value or a norm," and it refers to how similar measurement is to its agreed-upon value.
- **Specificity** measures the ability of a test to achieve a negative outcome for persons who do not have the disorder for which it is being measured (also known as the "true negative" rate).
- **Recall** or Sensitivity is the ability of a test to detect patients with a disease correctly. Sensitivity evaluates how much a test accurately produces a positive outcome (also known as the "true positive" rate) for those who have the disease for which it is being tested.
- **Precision** is described as "the performance of being exact" and relates to the proximity of two or more measurements, regardless of whether they are correct or not. Precision measurements have the potential to be inaccurate.

4. Results

Table I highlights the performance of the proposed BiLSTM model for cardiac arrhythmia detection versus varying BiLSTM layers. The number of filters for each layer is set to 32. In addition, Adam optimizer and binary cross-entropy loss are considered. From Table I, it can be observed that accuracy and specificity performance improve with increasing BiLSTM layers. In contrast, the recall performance slightly decreases with increasing BiLSTM layers. Another thing to notice is that the training performance in a majority of the cases is much higher than the validation performance. A possible option to improve the validation performance is to use a higher number of filters for each layer. Another option is to make use of nonlinear activation functions and batch normalization techniques. However, these changes come at the cost of increased complexity and latency.

Table 1: Performance of the proposed BiLSTM model for cardiac arrhythmia detection versus varying BiLSTM layers

BiLSTM layers	Module (Training/ Validation)	Accuracy (%)	Recall (%)	Precision (%)	Specificity (%)
1 Layer	Training	81.0	91.7	75.1	70.9
	Validation	56.9	40.5	39.9	66.0
2 Layers	Training	85.5	90.7	81.6	80.5
	Validation	55.7	30.1	35.8	69.9
3 Layers	Training	76.0	67.6	80.2	84.1
	Validation	63.6	35.3	48.8	79.4
4 Layers	Training	80.3	88.9	75.3	72.0
	Validation	56.0	11.6	25.2	80.8
5 Layers	Training	85.5	87.5	83.5	83.5
	Validation	56.1	67.0	18.7	83.6

5. Conclusion:

Prediction of cardiac arrhythmia through ECG signal analysis is gaining considerable importance to current researchers and clinicians. The BiLSTM model is one of the prominent deep learning models being used for arrhythmia detection. However, a study analyzing the impact of BiLSTM variation on the performance of cardiac arrhythmia detection is still lacking. In this paper, the performance of the BiLSTM model for arrhythmia classification with different numbers of BiLSTM layers is analyzed. It can be said that the accuracy and specificity performance of the model increases with increasing BiLSTM layers. As part of the future work, the impact of the number of filters, nonlinear activation functions, and dropout rate will be studied to achieve further performance improvements.

Acknowledgments:

The authors acknowledge the support of this research by the Yayasan Universiti Teknologi PETRONAS Fundamental Research Grant (YUTP-FRG) under Grant 015LC0-244.

References:

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:1801.02143.

George Moody, Roger Mark, (2005). MIT-BIH Arrhythmia Database. Retrived from https://www.physionet.org/content/mitdb/1.0.0/

Hassan, S. U., Zahid, M. S. M., & Husain, K. (2020, October). Performance comparison of CNN and LSTM algorithms for arrhythmia classification. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 223-228). IEEE.

He, R., Liu, Y., Wang, K., Zhao, N., Yuan, Y., Li, Q., & Zhang, H. (2019). Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. IEEE Access, 7, 102119-102135.

Hou, B., Yang, J., Wang, P., & Yan, R. (2019). LSTM-based auto-encoder model for ECG arrhythmias classification. IEEE Transactions on Instrumentation and Measurement, 69(4), 1232-1240.

Husain, K., Mohd Zahid, M. S., Ul Hassan, S., Hasbullah, S., & Mandala, S. (2021). Advances of ECG Sensors from Hardware, Software and Format Interoperability Perspectives. Electronics, 10(2), 105.

Lay-Ekuakille, A., Vergallo, P., Griffo, G., Conversano, F., Casciaro, S., Urooj, S., ... & Trabacca, A. (2013). Entropy index in quantitative EEG measurement for diagnosis accuracy. IEEE Transactions on Instrumentation and Measurement, 63(6), 1440-1450.

Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45-50.

Mukhopadhyay, S. K., Mitra, S., & Mitra, M. (2012). An ECG signal compression technique using ASCII character encoding. Measurement, 45(6), 1651-1660.

Oh, S. L., Ng, E. Y., San Tan, R., & Acharya, U. R. (2018). Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in biology and medicine, 102, 278-287.

Picon, A., Irusta, U., Álvarez-Gila, A., Aramendi, E., Alonso-Atienza, F., Figuera, C., ... & Eftestøl, T. (2019). Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia. PloS one, 14(5), e0216756.

Sharma, A., Garg, N., Patidar, S., San Tan, R., & Acharya, U. R. (2020). Automated pre-screening of arrhythmia using hybrid combination of Fourier–Bessel expansion and LSTM. Computers in Biology and Medicine, 120, 103753.

Xu, X., Jeong, S., & Li, J. (2020). Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. IEEE Access, 8, 125380-125388.

Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Computers in biology and medicine, 96, 189-202.

Detection of Cancer Cell and Tumor from MRI Image Using a Hybrid Approach – A Conceptual Framework

A F M Saifuddin Saif[a*], Zainal Rasyid Mahayuddin[b]

[a] Department of Computer Science, American International University - Bangladesh.
[b]Faculty of Information Science and Technology, University Kebangsaan Malaysia, Selangor, Malaysia
*Email: rashedcse25@yahoo.com

Abstract

Cancer is one of the deadliest diseases among all other diseases of human beings. Various types of methods for segmentation and detection of cancer cell or tumor were proposed by previous research. Among these methodologies of detection and segmentation, most of the previous methods can detect tumors or cancer cell accurately but have problems such as too complex computations for normal computers to process. In addition, some methods are unable to detect or draw out the region of the tumor with above 90% to 100% accuracy which is much needed for safe surgery. In this context, Magnetic Resonance Imaging (MRI) is an effective tool for cancer detection in recent years. This research proposes a conceptual framework for cancer cell and tumor detection from MRI images. For improved classification this research intends to use convolutional neural network. In addition, adaptive histogram equalization (CLAHE) is proposed to provide noise free images for improved performance of the overall methodology. Proposed conceptual framework is expected to provide robust solution as improved segmentation and detection of cancer cell and tumor from MRI images.

Keywords: Segmentation; Convolutional neural network; MRI images

1. Introduction

Cancer is a dense and abnormal cells proliferation in the body tissue. Cancer cells do not have the longer respond to some or many of the signals that mainly control cellular growth and death. Researchers proposed various methods for detecting cancer, i.e. Magnetic Resonance Imaging (MRI), microwave imaging, film-screen mammography. MRI is one of the most significant technology of cancer imaging for its accuracy. There are also some significant deep learning and machine-learning algorithms also proposed by the researchers, i.e. convolutional neural network (CNN) etc. This research proposes a conceptual framework where denoising approaches are considered and convolution neural network is intended to used for classification of tumor from input image. For pre-processing, this research intends to use adaptive thresholding for foreground extraction and adaptive histogram equalization method for image contrasting. The histogram equalization operation is considered to enhance the contrast of the image.

Rest of this paper is organized as follows. Section 2 presents core research background, section 3 demonstrates conceptual framework proposed by this research, section 4 illustrates proposed experimentation evaluation approach for the proposed conceptual framework, finally concluding remarks are presented in section 5.

2. Background Study

Cancer and tumor detection is a significant research in modern biomedical and computer vision research domain. Researchers proposed various methods for detection of cancer and tumor in modern technology, i.e. Microwave imaging, MRI, Film-screen mammography. MRI is mostly used for tumor detection. In this context,

researchers proposed various image processing and deep learning methods for cancer detection, segmentation and visualization in three-dimensional way.

A deep learning classification on MRI images for brain tumor detection and classification was proposed by Rathi and Palani (2015). They performed tumor classification using multiple kernel based probabilistic clustering and deep learning classifier. In the segmentation part, median filtering is used for image pre-processing and multiple kernel based probabilistic clustering. In feature extraction module, shape, texture, intensity based features are extracted. Linear Discriminant Analysis (LDA) was used for selecting important features. Deep learning classifier is used in classification module that is employed having two important processes of training phase and testing phase. However, their research requires further improvement towards robust experimentation. Müller et al. (2016) proposed an automatic brain tumor segmentation with deep neural network. Their proposed networks are tailored to glioblastomas pictured in MR images. To segment a brain, their proposed method requires between 25 seconds and 3 minutes that is one order of magnitude faster than most state-of-art methods. Convolutional Neural Network (CNN) was used to implement a novel two-pathway architecture that learns about the local details about the brain. However, their research requires further investigation towards lower computational complexity.

Damodharan and Raghavan (2015) proposed combined tissue segmentation and neural network for brain tumor detection where pre-processing part consists of skull stripping. The skull removed MRI images are employed for further classification of the brain tissues. In their research, initial steps involved in feature extraction which aimed to find the neighbor blocks of the entire divided blocks, finding distance between all the neighbor blocks, finding the feature values of the blocks with distinct distance measures. After the feature extraction by Damodharan and Raghavan (2015), MRI image classification using neural network starts using Feed Forward Neural Network (FFNN). However, Damodharan and Raghavan (2015) could concentrate more in denoising issues to improve overall performance.

3. Proposed Conceptual Framework

Conceptual framework proposed by this research consists of two main parts, i.e. preprocessing and segmentation part. Preprocessing part mainly contains noise reduction of MRI images and thresholding for foreground extraction. In this context, this research intends to use median filter and adaptive thresholding to denoise image in order to provide noise free image to the segmentation part. Another object of preprocessing is to enhance contrast of the images. In the preprocessing phase, this research intends to use binarization thresholding to extract whole head section in order to eradicate the background of the MRI images that may interfere with further processing. After this, adaptive histogram equalization is intended to be used to enhance the contrast of the images. However, in the context of MRI images, overall contrast is expected to be densely distributed in some portion of the image, not throughout the whole image. So, adaptive histogram equalization initializes a window for equalizing consisting pixels. Besides, this step is expected to provide clipping the excess contrast and redistribute among all histograms bins (Arahusky, 2021).

After preprocessing, maximum noise free image will be inputted in the convolutional neural network (CNN) where in this context this research is intended use a pretrained neural network. In the deep convolutional neural network that is intended to be used by this research, encoder part will be consisted of two convolutional layers. The first one is of stride 1 and the second one is of stride 2 followed by a max pool layer that uses nearest neighbor method for down-sampling which will be repeated two more times, each for lower resolutions. The decoder module will have other modules like encoder, i.e. max pool layer, an unpooling layer used for up-sampling.

4. Proposed Experimental Evaluation

This research intends to use accuracy (Mahayuddin & Saif , 2020; Saif et al. 2015;Schawkat et al., 2020), sensitivity (True Positive Rate) (Saif &Mahayuddin, 2020; Saif et al., 2015 ; Ide et al. ,2020) and specificity (True Negative Rate) (Saif et al., 2014 ;Beckett et al. ,2020) as primary performance metrics for segmenting and detecting tumor region from whole brain MRI image. All three metrics can be calculated using equation (1), (2) and (3).

$$Sensitivity = TP / (TP + FN) \quad (1)$$

$$Specificity = TN / (TN + FP) \quad (2)$$

$$Accuracy = (TN + TP) / (TN + TP + FN + FP) \quad (3)$$

True positive (TP) or tumor pixels marked as tumor which can be measured by checking how many of the ground truth tumorous pixels are marked as tumor pixel in the segmented tumor image. False positive (FP) or non-tumor pixels marked as tumor which can be measured by checking how many of the background pixels (non-tumor pixels) are marked as tumor pixels in the segmented tumor image. True negative (TN) or non-tumor pixels marked as non-tumor which can be measured by checking how many of the background pixels of the ground truth image are marked as background pixels in the segmented tumor image. False negative (FN) or non-tumor pixels marked as tumor. which can be measured by checking how many of the tumor pixels in the ground truth image are marked as background pixels in the segmented image. Sensitivity or true positive rate indicates the proportion of pixels of tumor region in MRI segmented correctly as tumor pixel. Specificity or true negative rate indicates the proportion of non-tumor pixels that are correctly segmented or marked by segmentation mask as non-tumor pixel. Accuracy is the rate of how much correctly the whole tumor is segmented, that is tumor marked as tumor and non-tumor marked as non-tumor together. For validation purpose, this research intends datasets from figshare (Divya et al. , 2020 ; Deepak and Ameer, 2020; Bulla et al., 2020) which contains 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumor: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices).

This research aims to use convolutional neural network for classification purpose after preprocessing step. In this context, MRI images are normally full of noises. Improved pre-processing of image data can be reduced as noise free sand prepared an image for next steps such as segmentation. In the case of segmentation, existing research methods reached over 70% to 80% accuracy, but for ensuring a safe surgery and other treatment method, tumor must be detected with 100% or as close to 100%. This research aims to establish an efficient segmentation method that can achieve an accuracy of almost 100% with as less complex computation method as possible.

5. Conclusion

Research on cancer cell and tumor detection and segmentation from MRI images is a valuable working area of modern computer vision research domain. Magnetic Resonance Imaging (MRI) is a medical imaging technique that is being used for cancer and tumor detection mostly nowadays. In this context, segmentation of medical imagery is a very challenging issue as MR images are full of many important information about patient's health and it is somewhat noisy to watch. This research proposed a conceptual framework to segment cancer cell from brain MRI image and later convolution network is considered for further processing. This research aims to use T1 MRI image dataset of brain tumor containing meningioma tumors and brain tumor segmentation (BraTS) challenge dataset of brain tumor. Later, experimental evaluation strategy is proposed to validate the proposed conceptual framework which is expected to contribute immensely in medical image processing domain. In the near future, proposed conceptual framework will be investigated extensively to establish the proposed methodology.

Acknowledgements

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the "Geran Universiti Penyelidikan" research grant, GUP-2020-064.

References

Arahusky.(2021).Tensorflow-Segmentation. Retrieved from https://github.com/arahusky/ Tensorflow-Segmentation.

Bulla, P., Anantha, L., & Peram, S. (2020). Deep Neural Networks with Transfer Learning Model for Brain Tumors Classification. *Traitement du Signal, 37*(4).

Beckett, A. J., Dadakova, T., Townsend, J., Huber, L., Park, S., & Feinberg, D. A. (2020). Comparison of BOLD and CBV using 3D EPI and 3D GRASE for cortical layer functional MRI at 7 T. *Magnetic resonance in medicine*, *84*(6), 3128-3145.

Damodharan, S., & Raghavan, D. (2015). Combining tissue segmentation and neural network for brain tumor detection. *International Arab Journal of Information Technology (IAJIT)*, 12(1).

Deepak, S., & Ameer, P. (2020). MSG-GAN Based Synthesis of Brain MRI with Meningioma for Data Augmentation.IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT).

Divya, S., Suresh, L. P., & John, A. (2020). A Deep Transfer Learning framework for Multi Class Brain Tumor Classification using MRI. 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).

Ide, M., Atsumi, T., Chakrabarty, M., Yaguchi, A., Umesawa, Y., Fukatsu, R., & Wada, M. (2020). Neural basis of extremely high temporal sensitivity: insights from a patient with autism. *Frontiers in Neuroscience*, *14*, 340.

Mahayuddin, Z. R., & Saif, A. S. (2020). A Comprehensive Review Towards Segmentation And Detection Of Cancer Cell And Tumor For Dynamic 3d Reconstruction. *Asia-Pacific Journal of Information Technology and Multimedia*, 9(1), 28-39.

Müller, S., Weickert, J., & Graf, N. (2016). Automatic brain tumor segmentation with a fast Mumford-Shah algorithm. Medical Imaging 2016: Image Processing.

Pizer, S. M. (1990). Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen *m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group.* First Conference on Visualization in Biomedical Computing, Atlanta, Georgia.

Rathi, V. G. P., & Palani, S. (2015). Brain tumor detection and classification using deep learning classifier on MRI images. *Research Journal of Applied Sciences, Engineering and Technology*, *10*(2), 177-187.

Saif, A. S., & Mahayuddin, Z. R. (2020). Vehicle Detection for Collision Avoidance Using Vision based Approach: A Constructive Review. *Solid State Technology, 63(2s)*, 2861-2869.

Saif, A. F. M. S., & Prabuwono, A. S. (2015). Moment Feature Based Fast Feature Extraction Algorithm for Moving Object Detection Using Aerial Images. *PLoS One*, *10*(6).

Saif, A. F. M. S., Mahayuddin, Z. R., & Prabuwono, A. S. (2015). Efficiency Measurement of Various Denoise Techniques for Moving Object Detection Using Aerial Images. International Conference on Electrical Engineering and Informatics (ICEEI).

Saif, A. F. M. S., Prabuwono, A. S., & Mahayuddin, Z. R. (2014). Moving object detection using dynamic motion modeling from UAV aerial images. *Scientific World Journal*, *2014*.

Schawkat, K., Ciritsis, A., von Ulmenstein, S., Honcharova-Biletska, H., Jüngst, C., Weber, A., . . . Reiner, C. S. (2020). Diagnostic accuracy of texture analysis and machine learning for quantification of liver fibrosis in MRI: correlation with MR elastography and histopathology. *European radiology*, *30*(8), 4675-4685.

Scheduling Strategies for Operating Room Surgical Scheduling Problems

Masri Ayob[a], Dewan Mahmuda Zaman[b]*

[a,b] *Data Mining & Optimization Research Group (DMO), Center for Artificial Intelligence (CAIT), Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia (UKM), Malaysia*
*Email:*mahmudazaman0509@gmail.com*

Abstract

Optimizing surgical scheduling problem (SSP) plays a crucial role in improving hospital health care services. Many studies tackled this problem with different perspectives, i.e. different solution approaches, decision levels, mathematical models and scheduling strategies. Thus, this work aims to review some literature in SSP that focused on scheduling strategies: open strategy, block strategy and modified block strategy. We discuss the strength and limitation of these scheduling strategies and pointed out the practicality of these strategies in real hospital practice. We also identified some approaches proposed in the literature related to each of these strategies. The study shows that block strategy is widely used in real life though it has some limitations, such as underutilization which can be tackled by applying modified block strategy. We also addressed that among several solution approaches, metaheuristics are frequently applied while considering different scheduling strategies.

Keywords: Surgical scheduling; Operating room; Scheduling strategy;

1. Introduction

In recent years, the operating room plays a crucial role in health care, and the demand for quality service is increasing day by day. The operating room is the core sector for both cost and revenue in a hospital (May et al, 2011). It is reported that more than 60% of patients are admitted for surgical operation, and undoubtedly, it is a very complex task to be tackled by the hospital management (Fügener et al,2017). The surgical scheduling problem (SSP) is a complex problem because it involves many resources and stages. Many works tackled this problem from different perspectives, such as solution approaches, decision levels, mathematical models and scheduling strategies. The targets of hospital managers are to maximize the utilization of operating room and this leads to different strategic plans (Gür & Eren, 2018). Thus, in this work, we focus on the view of scheduling strategies. Scheduling strategies can be grouped into three strategies: open strategy, block strategy and modified block strategy.

Block Scheduling Strategy: A scheduling strategy is called block strategy when each operating room is pre-allocated to a speciality with different surgery groups such as surgeons (Zhu et al, 2019). Each operating room can be divided into different time slots for different speciality, or can be allocated with a speciality for the whole day (Zhu et al, 2019). In block scheduling, some human resources such as nurses and anesthesiologist are assigned to each block, which is applicable for a cyclic timetable such as a certain number of weeks or months (Rahimi & Gandomi, 2020). This strategy can be classified into two types such as block strategy with master surgical scheduling and block strategy without master surgical scheduling (Gür & Eren, 2018).

Open Scheduling Strategy: This strategy is the opposite of the block scheduling strategy. There is no pre-allocation of any operating room. Surgery case can be assigned to any operating room regardless of the type (Agnetis et al, 2014). If the number of required resources is available, the surgery can be performed in any suitable operating room. Most of the time, open strategy follows First-Come-First-Serve principle (Zhu et al, 2019). Patients are scheduled without any speciality related restriction (Augusto et al, 2010). While assigning

patients to operating room, it is not necessary to check the speciality of the operating room. Any patient can be assigned to any available operating room.

Modified Block Scheduling Strategy: This is the combination of block and open strategy to use the advantages of both strategies. There are two ways to modify block strategy and apply it as modified block strategy (Fei et al, 2009). The first one is to reserve some operating room and to keep open for others. The second way is to release unused blocks for the remaining surgeries.

2. Advantage and Disadvantage of Different Scheduling Strategies

Each scheduling strategy has its advantages and disadvantages. Table 1 shows the comparison of these scheduling strategies.

Table 1. Advantages and disadvantages of different scheduling strategy

Strategy	Advantage	Disadvantage
Block Strategy	- In real life, block scheduling strategy is applied more often in many hospitals because the complexity of this strategy is lower than open strategy (Zhu et al, 2019). - Surgeons' preference is to centralize all the surgeries at a specific time of workday. So, applying this strategy is suitable for surgeons because working time is fixed in this strategy (Zhu et al, 2019). - This strategy reduces the time required for preparation and cleaning time between surgeries which also reduces the patient waiting time (Gür & Eren, 2018).	- In this strategy, the operating room is allocated to some specific surgeon, and if there is no surgical case for that surgeon on the day or the cancellation occurs, other surgeons cannot be assigned to that time block/operating room.
Open Strategy	- It is a more flexible strategy because there is no pre-allocated time block. - With the comparison of block scheduling strategy, the open strategy has better utilization of operating room (Zhu et al, 2019).	- A bad design of open system can cause a high loss rate. - This strategy is not popular in real life because it causes inconvenience, and also surgeons cannot centralize their work. - May have to face inconvenience when they have to deal with emergency cases because patients have priority to choose time (Zhu et al, 2019). - This strategy may cause long waiting time because of dynamic patient arrival (Zhu et al, 2019).
Modified Block Strategy	- This strategy is more convenient for operation room management because it can assign other operations to perform in the operation room to prevent the loss of late cancellations (Agnetis et al, 2014). - Capable to solve the problems mentioned for block and open strategy.	No disadvantage reported in the literature.

Although there are some limitations, block strategy is applied in many hospitals, for example, many hospitals in Europe adopt this strategy (Penn et al, 2017). In many hospitals, the block strategy is implemented due to surgeons' preferences (Zhu et al, 2019). Open strategy is flexible but may cause low utilization or delay. For all these reasons, open strategy is not mostly adopted in real life because this strategy is not popular among surgeons (Zhu et al, 2019). Many studies reported on block and open scheduling strategy but modified block

scheduling got least attention to study although it can be the solution of the limitations mentioned above. So here is an opportunity to investigate in future about modified block scheduling strategy.

3. Methods

There are many studies reported on different surgical strategy. Table-2 represents some studies on different scheduling strategy and methods used to solve the SSP in between the year 2016-2020.

Table 2. Studies on different scheduling strategies and their solving approaches

Paper	Block Strategy	Open Strategy	Modified Block Strategy	Solution Approach
Zhang et al. (2020)	✓			Column-generation-based heuristic
Khalfalli et al (2020)		✓		Adaptive Tabu Search
Zhu et al (2020)	✓			Hybrid Grey Wolf Optimizer - Variable Neighbourhood Search
Behmanesh et al (2019)		✓		Fuzzy Pareto envelope-based selection ant system
Lin & Chou (2019)		✓		Hybrid genetic algorithm
Khalfalli et al (2019)		✓		Tabu search
Moosavi et al (2018)	✓			MIP-based Local Search Neighborhood
Wu et al (2018)		✓		Hybrid genetic algorithm -variable neighborhood search
Nyman & Ripon (2018)		✓		Simulated Annealing, Genetic algorithm, Variable neighbourhood descent
Xiang (2017)		✓		Ant Colony Optimization
Guido & Conforti (2017)	✓			Hybrid genetic algorithm
Mateus et al (2017)	✓			Local search heuristics
Beroule et al (2016)		✓		Particle Swarm Optimization
Marchesi et al (2016)	✓			Genetic algorithm
Doulabi et al (2016)		✓		Branch-and-price-and-cut algorithm

Here, we took randomly three studies for five different years and tried to compare which scheduling strategy studied more on SSP. Table-2 represents that, in between the year 2016-2020, most of the studies are focused on open scheduling strategy and some are focused on block scheduling strategy. However, no studies are reported on modified block scheduling strategy. Table-2 also shows that metaheuristics such as tabu search, genetic algorithm, simulated annealing, particle swarm optimization etc. are mostly applied for solving both strategies. Other than the mention studies, (Liu et al, 2010) considered both open and block scheduling strategies and solved by heuristic approach. Although by applying modified block scheduling strategy, the drawbacks of other two strategies can be overcome, it is not studied enough.

3. Conclusion

Many works are done so far with different scheduling strategies but number of studies differs from year to year. For example, until 2010 block strategy got more attention. But now most of the studies focus on open scheduling. On the other hand, there is very few studies on modified block scheduling strategy. In real life block strategy is more preferable because human resources of surgery can centralize their work but still there is some limitations in this strategy. To overcome the problems of block strategy, modified block strategy can be applied which is also preferable for the management.

Acknowledgment

The authors wish to thank the Universiti Kebangsaan Malaysia and the Ministry of Higher Education Malaysia for supporting and funding this work (grant ID: TRGS/1/2019/UKM/01/4/1).

References

Agnetis, A., Coppi, A., Corsini, M., Dellino, G., & Meloni, C., Pranzo, M. (2014) A decomposition approach for thecombined master surgical schedule and surgical case assignment problems. Health Care Manag Sci 17(1):49–59

Augusto, V., Xie, X., & Perdomo, V. (2010) Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. Comput Ind Eng 58(2):231–238

Behmanesh, R., & Zandieh, M. (2019) Surgical case scheduling problem with fuzzy surgery time: An advanced bi-objective ant system approach. Knowledge-Based Systems:186

Beroule, B., Grunder, O., Barakat, O., Aujoulat, O., & Lustig, H. (2016, June). Particle Swarm Optimization for Operating Theater Scheduling Considering Medical Devices Sterilization. In International Conference on Swarm Intelligence Based Optimization (pp.41-56)

Doulabi, S. H. H., Rousseau, L. M., & Pesant, G. (2016a) A constraint-programming-based branch-and-price-and-cut approach for operating room planning and scheduling. INFORMS J Comput 28(3):432–448

Fei, H., Chu, C., & Meskens, N. (2009) Solving a tactical operating room planning problem by a column generation-based heuristic procedure with four criteria. Ann Oper Res 166(1):91–108

Fügener, A., Schiffels, S., & Kolisch, R. (2017) Overutilization and underutilization of operating rooms: insights from behavioral health care operations management. Health Care Manage Sci 20(1):115–128

Guido, R., & Conforti, D. (2017). A hybrid genetic approach for solving an integrated multi-objective operating room planning and scheduling problem. Computers & Operations Research Volume 87, Pages 270 282.

Gür, S. & Eren, T. (2018) Application of Operational Research Techniques in Operating Room Scheduling Problems: Literature Overview. Journal of Healthcare Engineering: Volume 2018

Khalfalli, M., Abdelaziz, F. B., & Kamoun, H. (2019). Multi-objective surgery scheduling integrating surgeon constraints. Management Decision

Khalfalli, M., Abdelaziz, F. B., Verny, J. & Masmoudi, M. (2020), Technology enhancement of surgeries scheduling: a bi-objective optimization model, Management Decision, Vol. 58 No. 11, pp. 2513-2525.

Lin, Y. K., & Chou, Y. Y. (2019). A hybrid genetic algorithm for operating room scheduling. Health Care Management Science, 1-15

Liu, Y., Chu, C. & Wang, K (2010) Aggregated state dynamic programming for operating theater planning. in Proceedings of the IEEE Conference on Automation Science and Engineering (CASE), pp. 1013–1018

Mateus, C., Marques, I., & Captivo, M. E. (2017) Local search heuristics ´ for a surgical case assignment problem, Operations Research for Health Care

Marchesi, J. F., & Pacheco, M. A. C. (2016) A genetic algorithm approach for the master surgical schedule problem. 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)

Moosavi, A, & Ebrahimnejad, S. (2018) Scheduling of elective patients considering upstream and downstream units and emergency demand using robust optimization. Computers & Industrial Engineering Volume 120

May, J.H., Spangler, W.E., Strum, D.P., & Vargas, L.G. (2011) The surgical scheduling problem: current research and future opportunities. Product Oper Manag 20(3):392–405

Nyman, J., & Ripon, K. S. N. (2018). Metaheuristics for the multiobjective surgery admission planning problem. In 2018 IEEE Congress on Evolutionary Computation (CEC)

Penn, M. L., Potts, C. N., & Harper, P. R. (2017) Multiple criteria mixed-integer programming for incorporating multiple factors into the development of master operating theatre timetables. Eur JOper Res 262(1):194–206.

Rahimi, I. & Gandomi, A. H. (2020) A Comprehensive Review and Analysis of Operating Room and Surgery Scheduling. Archives of Computational Methods in Engineering

Wu, X., Xiao, X., & Zhang, L. (2018) Optimizing the Three-stage Operating Room Scheduling Problem with RVNS-GA. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)

Xiang, W. (2017). A multi-objective ACO for operating room scheduling optimization. Natural Computing, 16(4), 607617

Zhang, J., Dridi, M., Moudni, A. E. (2020) Column-generation-based heuristic approaches to stochastic surgery scheduling with down stream capacity constraints. International Journal of Production Economics:Volume 229
Zhu, S., Fan, W., Yang, S., Pei, J., Pardalos, P.M., (2019). Operating room planning and surgical case scheduling: a review of literature. Journal of Combinatorial Optimization 37. 757–805
Zhu, S., Fan, W., & Liu, T. (2020) Dynamic three-stage operating room scheduling considering patient waiting time and surgical overtime costs. J Comb Optim 39**,** 185–215

Improving Production Rate and Growth Rate of Mutants: A Comparison of Constraint-Based Modeling Approaches

Kauthar Mohd Daud[a*], Zalmiyah Zakaria[b], Zuraini Ali Shah[c]

[a]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[b, c] Artificial Intelligence and Bioinformatics (AIBIG), School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
* Email: kauthar.md@ukm.edu.my

Abstract

In silico metabolic engineering is a process to improve the phenotypic characteristics by manipulating the genotypic elements of an organism. Reaction knockout is one of the genetic perturbation strategies use to analyse the effects in optimizing the production rate. The optimality of an organism can be predicted and simulated using constraint-based modeling (CBM) approaches. In this study, the stoichiometric model of *Escherichia coli* was examined to improve the production of succinic acid. Three well-known CBM approaches, which are Flux Balance Analysis (FBA), Minimization of Metabolic Adjustment (MoMA) and Regulatory On/Off Minimization (ROOM), were compared. The comparison resulted in a higher production rate predicted by MoMA and higher growth rate by FBA. Regardless, FBA is more stable in simulating the reaction knockout as it predicts the final steady-state of organisms after the genetic perturbation.

Keywords: constraint-based modeling method; reaction knockout; genome-scale metabolic network; in silico metabolic engineering

1. Introduction

In the early 1970s, the chemical industry has started to change into modern biotechnologies to fulfil the increasing demands of valuable products, for instance, pharmaceuticals, food ingredients and bio-based fuel (Maia, Rocha and Rocha, 2016). Advancements in genome sequencing have allow researchers to have insight on an organism. Therefore, the term "metabolic engineering" has coined that allow researchers to probe in detail the biological elements of an organism, thus gives full capability to exploit the genetic capability of an organism for strains improvement. The aims of metabolic engineering are to optimize the metabolism of organisms by exploiting and manipulating their metabolic capabilities through modelling. Thus, generates economically and industrially viable organisms through optimization and predictive tools.

In silico metabolic engineering is a scientific domain that integrates with the computational technology to allows faster simulation and design predictions of a mutant. It entails reconstructing the mathematical representation of the metabolic network, investigating the effects of genetic perturbations and their relations to the phenotype characteristics. Consequent to that, the development and reconstruction of genome-scale metabolic networks have rapidly grown and to keep up with the huge data, more sophisticated methods and algorithms are needed.

The approaches in metabolic engineering can be divided into two, which are the dynamic approach and static approach that varies in metabolic representation whereby dynamic approach uses kinetic modelling and static approach uses a stoichiometric matrix to represent the metabolic network (Suthers et al., 2021). Stoichiometric models are commonly used as modelling frameworks compared to kinetic models as it does not requires difficult-to-obtained experimental data for parameter estimation (Vasilakou et al., 2016). In stoichiometric models, the biochemical reactions in the metabolic network are represented as a set of stoichiometric equations, whereby the elements of different metabolites in the metabolic network are denoted as stoichiometric coefficients in the stoichiometric matrix.

1.1 Constraint-based Modeling Approaches

Organisms abide by the fundamental of evolution, whereby the fittest organism have more chances to survive than the less fit organism. In essence, a particular environment has specific characteristics, for instance, scarce resources, oxygen availability, and substrates presence. To be chosen for the next survival, the organisms must satisfy these constraints, thus limits the phenotypes. Therefore, an approach known as constraint-based modelling (CBM) has been developed. CBM is an approach to investigate the optimality of an organism by predicting and describing the metabolic phenotypes (Klamt et al., 2018). The feasible flux distributions space is created by constraining the systems, whereby only certain phenotypes are allowed to exist.

The unconstrained steady-state solution space is underdetermined due to the ratio of reactions typically exceeding the number of metabolites; thus a linear equation provides hyperplane that defines the allowable flux distributions. As a conclusion, the aim of constraint-based modelling (CBM) is to describe and predict the desired phenotypes of an organism by describing the metabolic networks of an organism using the stoichiometric framework and a series of constraints. Despite the imposition of constraints and steady-state assumption, the solutions generated are not limited to a single solution. Rather, the solutions generated are limited to the desired phenotypes.

Generally, there are three well-known approaches under constraint-based modelling methods - flux balance analysis (FBA), minimization of metabolic adjustment (MoMA), and regulatory on/off minimization (ROOM). Table 1 portrays the characteristics, advantages, and disadvantages of each constraint-based approaches as well as the applications that have been done.

Table 1. Summary of Constraint-based Approaches

Name	Characteristic(s)	Advantage(s)	Disadvantage(s)	Reference(s)
FBA	- Measure the optimal flux value of the desired objective function. - Linear programming.	- Enable analysis for large systems. - Suitable for linear and non-linear objective functions. - Able to predict lethality of a gene.	- Under certain medium/environmental conditions, the effects of regulatory constraints are not accounted. - Presence of multiple optima. - Not able to redesign the metabolic network.	- (Stalidzans et al., 2018) - (Budinich et al., 2017)
MoMA	Compare the steady-state fluxes after genetic perturbation between mutant and wild type. - Quadratic programming.	- Correctly predict the transient metabolic states.	- Only suitable for the new curate model that is not exposed to long-term evolutionary pressure. - The measured optimal flux is not a growth coupled.	- (Maia, Rocha and Rocha 2016)
ROOM	Minimize the number of significant flux changes between mutant and wild type. - MILP	- The predicted fluxes are nearer to the experimental data. - Favor for flux distributions that having high growth rates. - Able to predict lethality of a gene.	- Complex due to using binary variables in the objective function. - Able to find alternative shortest pathways, but these pathways are never being evolutionarily found by the organism.	- (Tomar and De, 2013)

2. Comparison of Constraint-based Approaches

In this study, FBA, MoMA and ROOM have been compared with *E.coli* model for optimizing the production of succinic acid. Previously, a total of 8 knocked out reactions resulted in higher performance in hybrid

algorithm, DSAFBA (Daud et al., 2018). Thus, the best production performance for 8 knocked out reactions predicted by FBA, MoMA and ROOM are depicted in the Fig. 2.

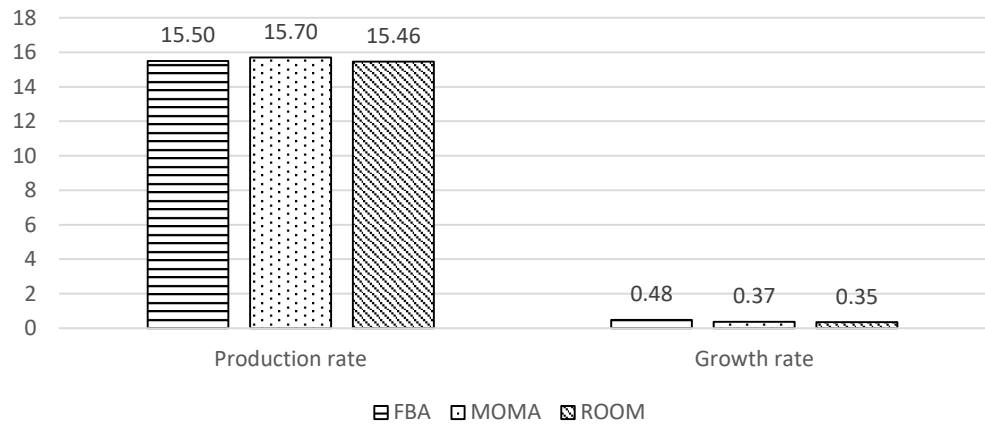


Fig. 2. Comparison of FBA, MoMA and ROOM.

The results showed that all three approaches could be well applied to produce metabolites, where MoMA exhibited a relatively better performance in terms of production rate. Meanwhile, FBA showed better performance compared to MoMA and ROOM in terms of growth rate. MoMA is able to find mutant with highest production rate but the production rate is evaluated during the intermediate state of organisms after genetic perturbations. On the contrary, FBA predicts the final steady-state of organisms after genetic perturbations.

After perturbations, the mutant will undergo slight or minimal redistribution before reaching a new steady state. MoMA is responsible to minimize the flux differences between mutant and wild type. However, MoMA only accurately identify fluxes at the early perturbations stage. Furthermore, this approach only applies to a new curated model that is not exposed to long-term evolutionary pressure. Meanwhile ROOM gives predictions nearer to the experimental data. However, it tends to modify and search for shortest pathways that gives maximum objective function with respect to the knockout. However, there is low chance that the alternative shortest pathways, are never being evolutionarily found by the organism.

After genetic manipulations, organisms evolve to a new steady-state activity patterns that satisfy constraints. FBA able to predicts the optimal long-term evolved state of the cell whereas ROOM and MoMA predicts the immediate initial outcome of genetic manipulations. Nevertheless, the cells will evolve from minimized flux distribution state to an FBA solution (Fong et al, 2005 and Doshi et al., 2020). In other words, genetic manipulations, first will lead to flux distribution predicted by MoMA and ROOM, then eventually converges to solution predicted by FBA.

3. Conclusion

Identifying respective reactions for knockout is a difficult and time-consuming process due to the complexity of metabolic models (Vasilakou et al., 2016). Furthermore, the complexity of models may lead to different combinations of reactions that producing different solutions. Hence, the solutions space become large. FBA is the most widely used method in assessing the model, although the other constraint-based approaches have been applied to the large metabolic models. This is because, FBA uses linear programming that is easier to apply than MoMA and ROOM which uses quadratic programming and mixed-integer linear programming, respectively. Though the solution provided by FBA is non-unique as it does not consider regulatory effects and metabolic concentrations, the existing metabolic networks are still incomplete, such as regulatory and kinetic

parameters (Maia, Rocha and Rocha 2016). Regardless of these imperfections, FBA can determine the steady-state fluxes of organisms as it does not require the above-mentioned parameters that are difficult to obtain.

References

Budinich, M., Bourdon, J., Larhlimi, A., & Eveillard, D. (2017). A multi-objective constraint-based approach or modeling genome-scale microbial ecosystems. PloS one, 12(2), e0171744.

Daud, K. M., Zakaria, Z., Shah, Z. A., Saberi, M., Mohamad, S. D., Omatu, S., & Corchado, J. M. (2018). A hybrid of Differential Search Algorithm and Flux Balance Analysis to Identify Knockout Strategies for in silico Optimization of Metabolites Production. Int. J. Advance Soft Compu. Appl, 10(2).

Doshi, P., Shri, M., Bhargava, P., Joshi, C. G., & Joshi, M. (2020). Microbial Production of Industrial Proteins and Enzymes Using Metabolic Engineering. In Engineering of Microbial Biosynthetic Pathways (pp. 189-204). Springer, Singapore.

Fong, S. S., & Palsson, B. Ø. (2004). Metabolic gene–deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nature genetics, 36(10), 1056-1058.

Klamt, S., Müller, S., Regensburger, G., & Zanghellini, J. (2018). A mathematical framework for yield (vs. rate) optimization in constraint-based modeling and applications in metabolic engineering. Metabolic engineering, 47, 153-169.

Maia, P., Rocha, M., & Rocha, I. (2016). In silico constraint-based strain optimization methods: the quest for optimal cell factories. Microbiology and Molecular Biology Reviews, 80(1), 45-67.

Suthers, P. F., Foster, C. J., Sarkar, D., Wang, L., & Maranas, C. D. (2020). Recent advances in constraint and machine learning-based metabolic modeling by leveraging stoichiometric balances, thermodynamic feasibility and kinetic law formalisms. Metabolic engineering, 63, 13-33.

Stalidzans, E., Seiman, A., Peebo, K., Komasilovs, V., & Pentjuss, A. (2018). Model-based metabolism design: constraints for kinetic and stoichiometric models. Biochemical Society Transactions, 46(2), 261-267.

Tomar, N., & De, R. K. (2013). Comparing methods for metabolic network analysis and an application to metabolic engineering. Gene, 521(1), 1-14.

Vasilakou, E., Machado, D., Theorell, A., Rocha, I., Nöh, K., Oldiges, M., & Wahl, S. A. (2016). Current state and challenges for dynamic metabolic modeling. Current opinion in microbiology, 33, 97-104.

Pembangunan Model Ramalan Tahap Risiko Pesakit Pembedahan Jantung Terbuka Menggunakan Pendekatan Pembelajaran Mesin

Development of a Risk Level Prediction Model for Open Heart Surgery Patients Using Machine Learning Approach

Norfazlina Jaffar @ Jaafar[a,b*], Afzan Adam[a*], Alwi Mohamed Yunus[b]

[a]Fakulti Teknologi dan Sains Maklumat Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia
[b]Institut Jantung Negara (IJN), 145 Jalan Tun Razak, 50400 Kuala Lumpur, Malaysia
*Email: norfazlina@ijn.com.my,afzan@ukm.edu.my

Abstrak

Penilaian klinikal awal terhadap pesakit yang akan menjalani pembedahan jantung terbuka merupakan satu proses penting bagi pengamal perubatan supaya dapat memilih dan memberikan rawatan yang terbaik sebelum, semasa dan selepas pembedahan. Oleh kerana masih tidak banyak kajian yang menggunakan kaedah pembelajaran mesin dalam menjangka hasil akhir kematian selepas menjalani pempedahan jantung terbuka, dan IJN hanya bergantung kepada model stratifikasi Bernstein-Parsonnet, EuroSCORE dan EuroSCORE II, model jangkaan ini dibangunkan dengan menggunakan kaedah pembelajaran mesin. Kajian ini menggunakan data pesakit yang tersedia daripada Registri Pembedahan Kardiotorasik dari tahun 2015 ke 2018 iaitu sebanyak 8787. Data-data ini telah menjalani kaedah pemprosesan awal data di mana ia merupakan salah satu proses penting dalam kaedah perlombongan data dan pembangunan model jangkaan. Terdapat 42 atribut yang dimasukkan ke dalam kajian, iaitu atribut-atribut yang digunakan dan telah di validasi oleh ketiga-tiga model yang digunakan di IJN. Pembangunan model stratifikasi risiko di IJN dijalankan dengan menggunakan regresi logistik, Artificial Neural Network (ANN), Random Forest (RF) dan Naïve Bayes. Data-data menjalani proses validasi dengan pecahan dataset kepada nisbah 60:40, 70:30 dan 80:20 untuk fasa latihan fasa pengujian. Kesemua hasil dapatan dibandingkan dengan menggunakan penilaian prestasi peratus kecekapan, AUC, kejituan, dapatan semula dan skor-F. Berdasarkan hasil dapatan kajian, model yang dibangunkan dengan menggunakan algoritma ANN dengan validasi pecahan dataset 80:20 mencapai prestasi yang terbaik iaitu dengan 82.74% ketepatan, 0.902 AUC, 0.8172 nilai kejituan, 0.8435 nilai dapatan semula dan skor-F 0.8301. Walaubagaimanapun, pembangunan model jangkaan ini perlu ditambahbaik dari segi pengumpulan dan penyediaan data, proses transformasi atribut dan mempelbagaikan lagi atribut-atribut yang relevan supaya memperoleh model jangkaan yang terbaik dengan penilaian prestasi yang tinggi.

Kata kunci: model jangkaan kematian;pembedahan jantung terbuka;kaedah pembelajaran mesin; pemprosesan awal data; *SMOTE*

1. Pengenalan

Sistem pengelasan tahap risiko atau dikenali sebagai sistem stratifikasi risiko, adalah model jangkaan skor risiko terhadap morbiditi dan kematian telah digunakan secara meluas di dalam bidang kardiovaskular dan pembedahan. Tahap risiko pesakit ditentukan dengan nilai pemberat atau nisbah yang diperolehi daripada model jangkaan di mana lebih tinggi nilai pemberat, lebih berisiko seorang pesakit berkenaan. Sistem Parsonnet adalah sistem yang pertama dibangunkan oleh Parsonnet V et.al pada tahun 1989. Pada tahun 1999, Nashef et.al memperbaharui dan membangunkan sistem European System for Cardiac Operative Risk Evaluation

(EuroSCORE) (Nashef et al., 1999). Namun begitu Garcia-Valentin et al. (2015), Gummert et al. (2009) dan Yap et al. (2005) yang telah menentusahkan model stratifikasi risiko EuroSCORE adalah terlebih jangkaan dan membuat kesimpulan model EuroSCORE tidak lagi sesuai digunakan dalam menentukan kadar risiko bagi pembedahan jantung pada masa kini. Beberapa penyelidik membuktikan model pembelajaran mesin adalah lebih tepat dalam menjangkakan kematian selepas pembedahan jantung terbuka dan mempunyai kelebihan yang terbaik berbanding EuroSCORE dan EuroSCORE II (Benedetto et al., 2020; Mejia et al., 2018; Kartal & Balaban, 2018; Allyn et al., 2017; Nouei et al., 2016). Benedetto, Dimagli, et al. (2020) melaporkan dalam kajian meta-analisis dan penilaian sistematik terhadap 15 penyelidikan (1997 – 2018), 10 (67%) melaporkan *Artifial Neural Network (ANN)* adalah model terbaik dengan kuasa diskriminasi yang terbaik, 2 (13.3%) melaporkan kaedah *ensemble*, 2 (13.3%) melaporkan *Decision Tree (DT)* atau *Random Forest (RF)* dan 1 (6.7%) melaporkan *Support Vector Machine (SVM)* sebagai model terbaik. Menurut Garcia-Valentin et al. (2015), Gummert et al. (2009) dan Yap et al. (2005) yang telah menjalankan kajian bagi menentusahkan model stratifikasi risiko EuroSCORE, model ini adalah terlebih jangkaan dan membuat kesimpulan model EuroSCORE tidak lagi sesuai digunakan dalam menentukan kadar risiko bagi pembedahan jantung pada masa kini. Salah seorang penyelidik tempatan telah menjalankan kajian bagi menentusahkan model EuroSCORE II terhadap pesakit yang menjalani pembedahan pintasan jantung koronari di IJN, mendapati nilai AUC bagi EuroSCORE II adalah 0.700 (95% SK 0.640 – 0.759) (Musa et al. 2018), di mana hanya menunjukkan kadar diskriminasi model ini di tahap baik sahaja. Menurut salah seorang penyelidik yang menjalankan kajian penilaian sistematik, terdapat hanya 8 kumpulan penyelidik atau institusi daripada 53 projek pembangunan model yang telah membangunkan model jangkaan kematian menggunakan teknik pembelajaran mesin iaitu *bootstrap bagging* (5) dan *Neural Network* (3) sehingga tahun 2017 (Karim et al. 2017). Ini menunjukkan masih tidak banyak kajian yang menggunakan kaedah pembelajaran mesin dalam menjangka hasil akhir kematian selepas menjalani pempedahan jantung terbuka. Permasalahan data atau kaedah pemprosesan awal data juga tidak banyak diterangkan di dalam penyelidikan-penyelidikan terdahulu. Permasalahan data hilang sering diabaikan oleh penyelidik dan hanya mengekalkan atribut yang tidak mempunyai data hilang di dalam analisis selanjutnya. Dengan pembuangan data hilang ini boleh menyebabkan peningkatan ketidaktepatan dalam menjangka hasil akhir dan juga boleh mengakibatkan kepada keputusan berat sebelah (*bias*).

1.1. Kaedah pemprosesan awal data

Kaedah pemprosesan awal data merupakan salah satu langkah kritikal sebelum proses menganalisis yang akan menghasilkan jangkaan yang lebih tepat, di mana kaedah ini di anggarkan melibatkan hampir 60% daripada keseluruhan proses (Taleb et al., 2015). Salah satu permasalahan data di dalam sektor kesihatan adalah data kelas tidak seimbang adalah keadaan dimana kelas yang ingin dijangkakan itu adalah jarang-jarang berlaku dan keadaannya adalah hanya minoriti berbanding kelas sebaliknya (Han et al., 2012). Chawla et al. (2002) telah memperkenalkan satu teknik bagi mengatasi masalah terkurang persampelan, iaitu dengan melakukan kaedah penambahan sampel sintetik dengan meniru data yang terdekat (*replicate*) yang menggunakan algoritma yang disebut sebagai *SMOTE (Synthetic Minority Over-sampling Technique)* dan menunjukkan prestasi yang superior dan signifikan terhadap data mamografi atau data klinikal yang mempunyai kelas tidak seimbang. Zhao et al. (2018) telah menggunakan teknik *SMOTE* yang digabungkan dengan regresi logistik bagi membangunkan kerangka pembelajaran terhadap data kesihatan yang mempunyai permasalah data tidak seimbang dan mendapati teknik ini menghasilkan keputusan jangkaan yang terbaik dengan peningkatan yang signifikan terhadap dapatan semula (*recall*) sebanyak 45.4%.

2. Metodologi Kajian

Data yang digunakan dalam kajian ini merupakan data berstruktur daripada tahun 2015 ke 2018 dan diperolehi daripada pengakalan data Registri Kardiotorasik IJN. Terdapat sejumlah 8787 data mentah yang merangkumi keseluruhan data pesakit yang menjalani pembedahan jantung terbuka di IJN, iaitu pembedahan pintasan koronari arteri (*Coronary Artery Bypass Graft, CABG*), pembedahan penggantian atau pemuliharaan injap (*Valve Replacement or Repair*) dan pembedahan yang melibatkan aorta. Daripada 49 atribut yang dipilih berdasarkan atribut-atribut yang terdapat di dalam model stratifikasi risiko Bernstein-Parsonnet, EuroSCORE dan EuroSCORE II, 3 atribut yang mempunyai data hilang melebihi 50% iaitu atribut *Planned Operation* (76.9%), *Planned Operation Aorta* (77.4%) dan *CL Pulmonary Hypertension* (54.8%) dikeluarkan daripada analisis selanjutnya disebabkan oleh peratus data hilang yang terlalu tinggi dan mungkin akan menyebabkan *bias* jika atribut-atribut ini menjalani proses penggantian data. Selain daripada itu, data latar belakang demografi pesakit yang diperoleh daripada sistem maklumat pesakit juga ditambah di dalam dataset mentah. Berdasarkan hasil penyelidikan terdahulu, kajian ini memilih untuk menggunakan perisian RapidMiner bagi pemprosesan awal data, *SMOTE*, pembangunan model regresi logistik, *Artificial Neural Network (ANN)*, *Random Forest (RF)* dan *Naïve Bayes* dan seterusnya penilaian prestasi model dengan menggunakan teknik validasi silang dengan pecahan dataset 60:40, 70:30 dan 80:20.

3. Dapatan Kajian

Berdasarkan penilaian prestasi setiap pecahan dataset, prestasi model jangkaan yang terbaik adalah model *Artificial Neural Network (ANN)* dimana semua nilai prestasinya adalah lebih tinggi berbanding Regresi Logistik, *Random Forest* dan *Naïve Bayes* di dalam setiap pecahan dataset. Jadual 1 menunjukkan penilaian prestasi bagi model ANN adalah yang tertinggi. ANN mempunyai kuasa diskriminasi yang lebih baik berbanding model regresi logistik, *Random Forest* (RF) dan *Naïve Bayes* apabila mempunyai lebih banyak maklumat atau lebih banyak data yang membolehkan algoritma ANN memjalankan proses latihan pembelajaran dan membuat jangkaan yang lebih tepat terhadap hasil akhir.

Jadual 1. Perbandingan Penilaian Prestasi Model-Model Jangkaan

Prestasi	*Logistic Regression*	*Artificial Neural Network*	*Random Forest*	*Naïve Bayes*
Ketepatan (%)	74.46%	**82.74%**	80.37%	72.75%
AUC	0.818	**0.902**	0.894	0.801
Kejituan	0.7390	**0.8172**	0.8180	0.7381
Dapatan Semula	0.7564	**0.8435**	0.7812	0.7054
Skor-F	0.7476	**0.8301**	0.7992	0.7213

4. Kesimpulan

Model yang dibangunkan menggunakan algoritma ANN yang mempunyai prestasi yang tertinggi berbanding algoritma lain dan model regresi logistik. Oleh itu, ANN merupakan algoritma pembelajaran mesin yang dicadangkan untuk membangunkan model jangkaan bagi pesakit yang menjalani pembedahan jantung terbuka. Kajian ini juga lebih memfokuskan kepada pemprosesan awal data berbanding kajian-kajian lain. Walaubagaimanapun, pembangunan model jangkaan ini perlu ditambahbaik dari segi pengumpulan dan penyediaan data, proses transformasi atribut dan mempelbagaikan lagi atribut-atribut yang relevan.

Pengenalpastian faktor-faktor risiko ini perlu dijalankan dengan lebih meluas dan *robust* supaya hasil dapatan kajian ini akan memperoleh pengetahuan baru atau *Knowledge Discovery of Database (KDD)*.

Penghargaan

Setinggi-tinggi perhargaan kepada semua pensyarah program Sains Data di Fakulti Teknologi dan Sains Maklumat UKM, pakar-pakar dan pegawai perubatan di Jabatan Pembedahan Kardiotorasik IJN, Pengarah, Ketua Jabatan dan ahli Unit Pengurusan Data dan Bantuan Biostatistik di Jabatan Penyelidikan Klinikal IJN, atas tunjuk ajar dan kerjasama yang diberikan semasa kajian ini dijalankan.

Rujukan

Allyn, J., Allou, N., Augustin, P., Philip, I., Martinet, O., Belghiti, M., Provenchere, S., Montravers, P., & Ferdynus, C. (2017). A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis. *PLoS ONE*, *12*(1), 1–12. https://doi.org/10.1371/journal.pone.0169772

Benedetto, U., Dimagli, A., Sinha, S., Cocomello, L., Gibbison, B., Caputo, M., Gaunt, T., Lyon, M., Holmes, C., & Angelini, G. D. (2020). Machine learning improves mortality risk prediction after cardiac surgery : Systematic review and meta-analysis. *The Journal of Thoracic and Cardiovascular Surgery*. https://doi.org/10.1016/j.jtcvs.2020.07.105

Benedetto, U., Sinha, S., Lyon, M., Dimagli, A., Angelini, G., & Sterne, J. (2020). *Can machine learning improve mortality prediction following cardiac surgery ? 0*(May), 1–7. https://doi.org/10.1093/ejcts/ezaa229

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(2), 321–357.

Garcia-Valentin, A., Mestres, C. A., Bernabeu, E., Bahamonde, J. A., Martín, I., Rueda, C., Domenech, A., Valencia, J., Fletcher, D., Machado, F., & Amores, J. (2015). Validation and quality measurements for EuroSCORE and EuroSCORE II in the Spanish cardiac surgical population: A prospective, multicentre study. *European Journal of Cardio-Thoracic Surgery*, *49*(2), 399–405. https://doi.org/10.1093/ejcts/ezv090

Gummert, J. F., Funkat, A., Osswald, B., Beckmann, A., Schiller, W., Krian, A., Beyersdorf, F., Haverich, A., & Cremer, J. (2009). EuroSCORE overestimates the risk of cardiac surgery: Results from the national registry of the german society of thoracic and cardiovascular surgery. *Clinical Research in Cardiology*, *98*(6), 363–369. https://doi.org/10.1007/s00392-009-0010-8

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Data mining concepts and techniques. In *Morgan Kaufmann, Elsevier*.

Karim, M. N., Reid, C. M., Cochrane, A., Tran, L., Alramadan, M., Hossain, M. N. & Billah, B. 2017. *Mortality risk prediction models for coronary artery bypass graft surgery: Current scenario and future direction. Journal of Cardiovascular Surgery* 58(6): 931–942. doi:10.23736/S0021-9509.17.09965-7

Kartal, El., & Balaban, M. E. (2018). *Machine learning techniques in cardiac risk assessment*. *26*(3), 394–401. https://doi.org/10.5606/tgkdc.dergisi.2018.15559

Mejia, O. A. V., Antunes, M. J., Goncharov, M., Dallan, L. R. P., Veronese, E., Lapenna, G. A., Lisboa, L. A. F., Dallan, L. A. O., Pomerantzeff, P. M. A., Brandão, C. M. A., Tarasoutchi, F., Zubelli, J., & Jatene, B. (2018). *Predictive performance of six mortality risk scores and the development of a novel model in a prospective cohort of patients undergoing valve surgery secondary to rheumatic fever*. 1–14.

Musa, A. F., Cheong, X. P., Dillon, J. & Nordin, R. Bin. 2018. *Validation of EuroSCORE II in patients undergoing coronary artery bypass grafting (CABG) surgery at the National Heart Institute, Kuala Lumpur:*

a retrospective review. F1000Research 7(May): 534. doi:10.12688/f1000research.14760.1
Nashef, S. A. M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., & Salamon, R. (1999). *European system for cardiac operative risk evaluation (Euro SCORE)*. *16*, 0–4.
Nouei, M. T., Kamyad, A. V., Sarzaeem, M. R., & Ghazalbash, S. (2016). Fuzzy risk assessment of mortality after coronary surgery using combination of adaptive neuro-fuzzy inference system and K-means clustering. *Expert Systems*, *33*(3), 230–238. https://doi.org/10.1111/exsy.12145
Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big Data Pre-processing: A Quality Framework. *Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015*, 191–198. https://doi.org/10.1109/BigDataCongress.2015.35
Yap, C. H., Mohajeri, M., Ihle, B. U., Wilson, A. C., Goyal, S., & Yii, M. (2005). Validation of EuroSCORE model in an Australian patient population. *ANZ Journal of Surgery*, *75*(7), 508–512. https://doi.org/10.1111/j.1445-2197.2005.03440.x
Zhao, Y., Wong, Z. S. Y., & Tsui, K. L. (2018). A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. *Journal of Healthcare Engineering*, *2018*(2010), 6275435. https://doi.org/10.1155/2018/6275435

Cybersecurity Threats and Practices in Internet Café: An Assessment of Cybercafé in Nigeria

Mansur Aliyu[a*], A. S. Baiti[b], A. B. Tambuwal[c], Samaila Musa[d], Aminu Aliyu[e]

[a]Sokoto State University, Sokoto, Nigeria
[b,d,e]Umaru Ali Shinkafi Polytechnic, Sokoto, Nigeria
[c]Federal University Gusau, Zamfara, Nigeria
* Email:mansur.aliyu@ssu.edu.ng, mansuraliyu@gmail.com

Abstract

The Internet has grown over the years from data transmission media to global cyberspace providing access to information to all kinds of users. Internet Users had gone beyond sending/receiving office related work to conducting personal financial transactions, socializing with family and friends, and posting the tiniest details of their private life. With these kinds of information and the billions of users having access to the Internet, a high cybercrime rate is inevitable, especially in developing countries like Nigeria. This paper attempts to explore cybersecurity threats and practices in Internet cafés in selected states in Nigeria. A survey method was adopted to administer questionnaires directly to cyber café operators and users in Sokoto and Kebbi states. A total of sixty (60) Internet cafés and computer business centers were surveyed using both close-ended and open-ended questions related to their current cybersecurity threats and practices. The significant findings revealed that most cybercafé operators in the two states had knowledge and experience about cybercrime. They all acknowledge the risks of cybercrime, but they took insufficient measures to avert unauthorized access to their computers. The findings also revealed that most operators use weak administrator and user passwords to log in to their networks, with no strong firewalls and up-to-date antivirus software, which resulted in corrupting computer files, data hacking, personal and financial identities theft, denial of service, virus infection, etc. In the end, it was recommended that both the cybercafé operators, general users, and the government take severe steps in securing their data, information, and network infrastructure against cybercrimes or threats.

Keywords: Cybersecurity; cybercrime; cybercafe; Internet; Nigeria

1. Introduction

Internet usage has become part of everyday life and has fundamentally changed our habits regarding data and information communication and processing. Nowadays, technological developments in developing countries have the potential to grow economic, social, and political changes. They also can advance criminal activities in any given country. As a nation, Nigeria has not been left behind in internet penetration and usage, primarily via mobile phones. It is a country prone to attacks by cybercriminals and a possible source of cybercrime activities (Makare, 2017). While most cybercrime attacks might target financial institutions like banks, internet users within the general public are also likely to become victims of similar criminal activities (Kshetri, 2019). Therefore, it is essential to assess both the awareness and preparedness level of local internet operators and users to deal with the threats of cyber-criminal activities.

Cybercrime refers to any criminal activity executed through the Internet (Osho & Adepoju, 2016; Aneke, et al. 2020). This involves many things from denial of service, downloading illegal files, non-delivery of goods or services and computer intrusions (hacking) to intellectual property rights abuses, economic espionage (theft of trade secrets), online extortion, international money laundering, identity theft, and a growing list of other Internet-facilitated offenses (Ajeet, 2014). Cybercrime is most difficult to immediately detect the method used to carry out the Crime, to know precisely where and when the users carried out the Crime. The anonymity of

the Internet makes it an ideal channel and instrument for many organized criminal activities (Ajeet, 2014; Omodunbi et al. 2016). The speed of cyber technology changes always beats security agencies' efforts, making it difficult for them to identify the origin of cybercrimes (Majesty, 2010; Roshan, 2008). As such, cybercafé operators and system developers need to consider developing an in-built tracking system that can detect and block all suspicious activities on their servers before the intrusion (Aliyu et al., 2020).

In developing countries like Nigeria, Internet café (i.e., cybercafé), also called business centers, are roadside shops opened specifically to provide computer-related services such as Internet access, typesetting, photocopy, scanning, printing, hardware repairs, software installation, computer training, etc. (Sodiq, 2012). The Internet cafés provide great commercial opportunities to small-scale entrepreneurs and employment sources to the youth, ultimately supporting the developing economy. It is crucial to protect these small scale businesses' IT infrastructure and assets from cyber threats and crimes. Since Internet café are not as buoyant as other institutions like banks, hospitals, agencies, examinations bodies, etc. According to Frank and Odunayo (2013), the cybercrime phenomenon has become a sophisticated and extraordinary increase recently and therefore called for a quick response in providing laws as highlighted by Ezeanokwasa (2019) that would protect cyberspace and its users. Cybercrime is involved and committed mostly from remote locations, making it difficult for police. The absence of enabling regulation makes policing even more difficult. Statistically, Nigeria ranked 43 in EMEA and ranked third among the ten nations that commit cybercrime in the world (Frank & Odunayo, 2013; Adebusuyi, 2008) Even though, the National Cybersecurity Initiatives (NCI) that was created in 2003 are yet to meet the proposed desired objectives, despite the help from the Nigerian cybercrime working group (NCWG) (Awhefeada & Bernice, 2020). Therefore, the government through the ministry of communication and digital economy must come in to protect the private sector, IT infrastructures, and facilities for information security and economic development.

Presently, cybercafés provide services such as personal browsing, emails, filling application forms (i.e., for jobs, admissions, exams, visas, licensing, etc.), online exams (CBT), academic research, online video games, and entertainment, etc. In higher institutions, cybercafés are the hub of accessing the Internet for assignments and final year projects. Unfortunately, while Internet cafes helped enhance IT adoption in the country, they have also allowed multiplying its abuses. Some youth visits cybercafés to access pornographic materials. Despite the consistent fight against Internet pornography in the country, only a few cybercafés where content filters are downloaded and installed to filter unwanted Internet content (Kshetri, 2019; Geoff et al. 2005; Longe et al., 2005). Despite majority of the cybercafés placed warning notices against surfing pornographic sites and spamming activities, still many users often ignored the notice, and keep sending spam mails, browse sex sites, surf and download unauthorized contents (such as video films, musical audios and other multimedia contents). Apart from the readiness and usage of Internet facilities in cybercafés for pornography and other cybercrimes, the installation of fixed wireless facilities in the Nigerian network landscape has added another dimension to the cybercrimes problem (Longe et al., 2005).

Similarly, the yahoo boys used cybercafés across the country as a medium for safe criminal activities against vulnerable users. They are engaged in illegal activities such as hacking e-commerce sites, bank accounts, ATM cards, email accounts, examination systems, travel sites, etc. As a result, phishing has become very popular as criminals simulate product websites to deceive innocent Internet users into submitting their financial credentials while ordering fake products (Longe et al., 2005). Thus, many cybercafés have been sealed off by security agencies due to the perpetration of cybercrimes, e.g., spamming, credit card fraud, ATM frauds, phishing, and identity theft using that café network (Olumide and Victor, 2010; Augustine, 2010). Currently, there is a lack of standardized up-to-date cybersecurity guidelines on the establishment and operations of cybercafé. Some cybercafés were shut down due to a lack of patronage by people scared of scammed, hacking, or virus attacks.

This study's primary purpose is to investigate and assess the current cybersecurity threats and practices affecting Internet café operations in Sokoto and Kebbi States of Nigeria; the paper explored the current cybersecurity countermeasures adopted by the selected Internet Café. Specifically, the paper examines the level of cybercrime and security awareness among operators, their knowledge about cybercrime and security threats,

and their experience on cybercrimes incidences that occurred in their cafés. Lastly, computer security scenarios were posed to them to assess their level of understanding of cybersecurity.

2. Research Methodology

To achieve this study's objectives, a survey design was used to administer questionnaires directly to the respondents at their respective Internet Café by the Researcher. Respondents' consent was sort utilizing a cover letter before filling the questionnaire. The letter mentioned the study's purpose, and with contact information of the Researcher if the respondent needs further clarification or additional explanation where necessary.

The introductory part of this paper begins with the study of the existing literature on cybersecurity threats and practices affecting Internet Café operations in Sokoto and Kebbi States. Books, academic papers, journals and were very significant data sources in this preliminary study. Thereafter, a questionnaire was designed from the information gathered through the literature. A survey using a self-administered questionnaire was carried out to collect the primary data from the Internet operators based on randomly selected cyber cafés and some computer business centers that are rendering the Internet services. The research results are supposed to be an accurate reflection of the target population; thus, care was taken when sampling to ensure the validity and reliability of the data collected. To minimize bias during data collection, a multi-stage random sample was done on the target population. First, select a cybercafé at random, then randomly choose the established cybercafé's respondents. The data collected was analyzed using SPSS v17 tool.

The Table 1 results of the respondent's demographic factors indicates that there were more male cybercafé operators in Sokoto and Kebbi states having 75.0% and 76.2% as compared to female with 25.0% and 23.8% respectively. The results also showed that most people operating/managing Internet cafés in Sokoto and Kebbi are young people below 33. They account for over 89.3% in Sokoto and 76.2% in Kebbi state. On educational qualification, the results indicate that most cybercafé operators in Sokoto State had a Diploma (39.3%) while in Kebbi state had Bachelor/HND (47.6%). It is expected that their higher qualification will enable them to have a better understanding of cybercrimes. It also shows that 82.1% and 76.2% of the operators in Sokoto and Kebbi had undergone computer training, workshop, or seminar, respectively. Moreover, 85.7% of cybercafé operators in Sokoto are aware of cybercrime, and 81.0% of the operators in Kebbi were mindful of the existence of the cybercrime. Lastly, Table I reveals that respondents from Sokoto and Kebbi metropolises often heard people talking or discussing cybercrime issues with 42.8% from Sokoto and 61.9% from Kebbi.

Table 1. Demographic Profile

Demographic	Description	Sokoto		Kebbi	
		N	%	N	**%**
Gender	Male	21	75	16	76.2
	Female	7	25	5	23.8
Age	18 – 22	11	39.3	6	28.6
	23 – 27	10	35.7	6	28.6
	28 – 32	4	14.3	4	19.0
	33 – 37	1	3.6	1	4.8
	38 and above	2	7.1	4	19.0
Qualification	Postgraduate	3	10.7	4	19.1
	Bachelor/HND	8	28.6	10	47.6
	Diploma	11	39.3	2	9.5
	SSCE	6	21.4	5	23.8
Computer Training	No. of Trained Cybercafé Operators	23	82.1	16	76.2
		5	17.9	5	23.8
Cybercrime Awareness	Operators Cybercrime Awareness	24	85.7	17	81.0
		4	14.3	4	19.0
Cybercrime Frequency	Very Often	11	39.3	5	23.8
	Often	12	42.8	13	61.9
	Rarely	4	14.3	1	4.8
	Never	1	3.6	2	9.5

3. Analysis of Findings

To find out whether the cyber café operators have any knowledge about the existence of cybercrimes, two questions in different sections of the questionnaire were posed to them. The first one in section B of the questionnaire wanted to find out whether, as a cybercafé operator, he/she is aware of the existence of cybercrimes. The second question was posed in section C, and it required the respondent to rate himself/herself on a Likert scale of 1 to 5 on the agreement with the statement that "there are risks involved when I am working online." On getting the two questions' descriptive statistics, it shows that most cyber café operators in both Sokoto and Kebbi acknowledge that there are risks while working online. At the same time majority of the respondent, port to be aware of computer crimes. Despite being knowledgeable about computer crime, the operators neglected to take active measures to avoid cybercrime and the high risk involved when working online.

Table 2, under the Sokoto section, describes how the cyber café operators in Sokoto know about cybercrime. 89.2% of respondents agreed that there are risks involved whenever they work online, 7.2% disagreed, and 3.6% were neutral. 78.6% of respondents agreed that they were aware of the various computer cybercrimes likely to be exposed to while working online, and only 3% are neutral and disagreed response. Also, 78.6% of respondents strongly agreed that it is always advisable to log in as a user rather than an administrator whenever going online, 3.6% agreed, and 17.8% were neutral. 46.4% of respondents were agreed that there is no risk in using the same password for different accounts or computers; 7.2% are neutral, while 46.4% disagreed. 53.6% of respondents agreed that they needed to regularly change their password/frequently to avoid cybercrimes, while 35.7% disagreed, 10.7% were neutral. Lastly, 89.2% of respondents agreed that any password they use should have at least eight characters, a combination of alphabets, digits & symbols, and 3.6% each were neutral and disagreed.

The Kebbi section of Table 2 indicates that 90.4% of the respondents agreed that there are risks involved whenever they work online, while 4.8% were disagreed and neutral. 85.7% of respondents agreed that they were aware of the various computer cybercrimes likely to be exposed to while working online, 9.5% were neutral, and 4.8% have disagreed. Also, 80.9% of respondents agreed that it is always advisable to log in as users rather than an administrator whenever going online, 4.8% disagreed, and 14.3% were neutral. 38.1% of respondents agreed that there is no risk using the same password for different accounts or computers; 33.3% disagreed 28.6% were neutral. 66.7% of respondents decided to regularly change their password/frequently to avoid cybercrimes, while 4.7% disagreed, and 28.6% were neutral. Lastly, 90.5% of respondents agreed that any password they use should have at least eight characters, which should be a combination of alphabets, digits & symbols, and 9.5% of respondents were neutral. Therefore, it shows that most cyber café operators from Sokoto and Kebbi know about cybercrime, but they take minimum protection of their computer systems from cybercrime.

Table 2. Knowledge about Cybercrimes from Cyber Café operators

State	Statement	Statement No.	Number of Respondent/Percentage (%)		
			Agree (%)	Neutral (%)	Disagree (%)
Sokoto	There are risks involved whenever I am working online.	Q1	25 (89.2)	1 (3.6)	2 (7.2)
	I am aware of the various cybercrimes I am likely to be exposed to while working online.	Q2	22 (78.6)	3 (10.7)	3 (10.7)
	It is always advisable to log in as a user rather than an administrator whenever going online.	Q3	22(78.6)	5 (17.8)	1 (3.6)
	There is no risk in using the same password for different accounts.	Q4	13 (46.4)	2 (7.2)	13 (46.4)
	I need to change my password regularly/frequently.	Q5	15 (53.6)	3 (10.7)	10 (35.7)
	Any password I use should have at least eight characters, a combination of alphabets, digits & symbols.	Q6	25 (89.2)	1 (3.6)	2 (7.2)
Kebbi	There are risks involved whenever I am working online.	Q1	19 (90.4)	1 (4.8)	1 (4.8)
	I am aware of the various computer crimes I am likely to	Q2	18 (85.7)	2 (9.5)	1 (4.8)

be exposed to while working online.				
It is always advisable to log in as a user rather than an administrator whenever going online.	Q3	17(80.9)	3 (14.3)	1 (4.8)
There is no risk in using the same password for different accounts.	Q4	8 (38.1)	6 (28.6)	7 (33.3)
I need to change my password regularly/frequently.	Q5	14 (66.7)	6 (28.6)	1 (4.7)
Any password I use should have at least eight characters, a combination of alphabets, digits & symbols.	Q6	19 (90.5)	2 (9.5)	0 (0)

Table 3 under the Sokoto section, describes the computer security scenarios viewed by cybercafé operators in the Sokoto metropolis. 96.4% of respondents were likely to have their computer system becoming corrupted by a virus, and 3.6% said it is unlikely. 57.2% said it is expected that a hacker took over their computers, 28.5% says it is doubtful, and 14.3% were neutral. Also, 85.8% of respondents were likely to become corrupted by a computer virus, 28.6%, 7.1% were neutral and very unlikely. 57.2% of respondents were very likely that personal identity being stolen, 25% were neutral, and 17.8% were unlikely. 42.9% of respondents were likely that personal/financial identities being stolen, 17.9%, 42.8% were doubtful, and 14.3% were neutral. 50% of respondents were likely not able to access the Internet because of a computer virus attack, 42.9% were unlikely, and 7.1% were neutral

Table 3 under the Kebbi section describes the computer security scenarios viewed by cyber café operators in the Sokoto metropolis. 95.2% of respondents were likely to have their computer system becoming corrupted by a virus, and 4.8% were neutral. 52.4% of respondents were possible that a hacker took over their computers, 19.0% were unlikely, and 28.6% were neutral. 95.2% of respondents were likely to have their files becoming corrupted by a computer virus, and 4.8% were doubtful. 76.2% of the respondents were alleged to have personal identity stolen, 4.8% were unlikely, and 19.0% were neutral. Under the statement that said personal/financial identities being stolen (loan fraud), 47.6% of respondents were likely, 28.6% were neutral, and 23.8% were unlikely. 61.9% of respondents agreed that they were likely unable to access the Internet because of a computer virus attack, and 38.1% were unlikely. 61.9% of respondents were probably having a computer infected with a virus due to visiting a website, 14.3% were dubious, and 23.8% were neutral. It, therefore, shows that both Sokoto and Kebbi respondents have a low-level adoption of computer security majors and mostly have systems corrupted by different kinds of viruses.

Table 3. Computer Security Scenarios from Cyber Café operators

State	Statement	No. of Respondents	Number of Respondent/Percentage (%)		
			Likely (%)	Neutral (%)	Unlikely (%)
Sokoto	My computer system became corrupted by a virus.	Q1	27(96.4)	0 (0)	1 (3.6)
	My computer being taken over by a hacker.	Q2	16(57.2)	4 (14.3)	8 (28.5)
	My files became corrupted by a computer virus.	Q3	24(85.8)	2 (7.1)	2 (7.1)
	My identity being stolen (credit cards, social security numbers, etc.).	Q4	16 (57.2)	7 (25)	5 (17.8)
	My financial identity being stolen (ATM PIN, Card number, secret code, etc.).	Q5	12(42.9)	4 (14.3)	12 (42.8)
	I am not able to access the Internet due to a computer virus.	Q6	14 (50)	2 (7.1)	12 (42.9)
	My computer became infected with a virus as a result of visiting a website.	Q7	16 (57.1)	3 (10.7)	9 (32.2)
Kebbi	My computer system became corrupted by a virus.	Q1	20 (95.2)	1(4.8)	0(0)
	My computer being taken over by a hacker.	Q2	11 (52.4)	6(28.6)	4(19.0)
	My files became corrupted by a computer virus.	Q3	20 (95.2)	0(0)	1(4.8)
	My identity being stolen (credit cards, social security numbers, etc.).	Q4	16 (76.2)	4(19.0)	1(4.8)
	My financial identity being stolen (ATM PIN, Card number, secret code, etc.).	Q5	10 (47.6)	6(28.6)	5(23.8)
	I am not able to access the Internet due to a computer virus.	Q6	13 (61.9)	0 (0)	8(38.)

	My computer became infected with a virus as a result of visiting a website.	Q7	13 (61.9)	5(23.8)	3(14.3)

Table 4, the spam mail incidences that are experienced by cyber café operators within the year. It shows below, among the incidences that experienced within the cybercafé in Sokoto and Kebbi, after grouping the incidences, 12 of the 35 respondents in Kebbi indicated that they had experienced spam mail than cybercafé. Only nine respondents in Sokoto experienced spam mails, among other incidences. There is also an indication of denial of service, theft of computer services, and virus or worm attack incidences experienced in cyber café in Sokoto than Kebbi.

Table 4. Factors of Cybercrime experienced in Sokoto and Kebbi States

S/No.	Cyber Crime Incidences Experienced	No. of Respondents in Sokoto	No. of Respondents in Kebbi
1.	Spam mail	9	12
2.	Denial of service	7	3
3.	Hacking	5	2
4.	Identity theft	2	1
5.	Computer crashing	5	2
6.	Theft of computer services	7	5
7.	Virus or worm attack	7	6
8.	Trojan or root-kit attack	2	4
	Total	44	35

5. Conclusion

The research findings revealed that there is knowledge about computer cybercrimes among the two States operator, out of which only 22(78.6%) and 18(85.7&) respondents from Sokoto and Kebbi respectively were aware of the various computer crimes exposed to while working online. Due to minimum protection or negligence on the cyber threats, there are high risks of attacks in most cybercafés across the two states (Kebbi with 90.4% and Sokoto with 89.2%). Both states experience high rates of computer virus attacks on their hardware, software installations, and files and data. Cybercrime activities need to be checked and quickly addressed as they affect small-scale businesses (cybercafé owners) and the general public. Addressing the menace of cybercrime should involve a holistic approach from all stakeholders – users, operators, internet service providers, cybersecurity agencies, and the government. Therefore, there is a need to begin public enlightenment campaigns on cybersecurity and best practices, provide adequate and affordable tools to prevent cyber-attacks and information theft, create an enabling environment for the operators, and create appropriate laws to combat cybercrime perpetrators.

Further measures shall be put in place to enhance preparedness for handling risks of computer crimes among the general public. The availability and price of antivirus and other software meant to enhance the preparedness levels should be made accessible. This way, it will be easier for the general public members to prepare adequately to deal with threats of computer crimes. Educating the general public on how they should safeguard their information when going online should also be undertaken regularly. It would be essential to carry out group discussions with the internet managers and end-users to determine the extent of awareness and preparedness about computer crimes; this is rather than relying only on the self-administered questionnaire to assess the level of understanding and preparedness. This way, issues that are likely to be confusing are sorted out and clarified in data collection.

References

Adebusuyi, A. (2008): The Internet and Emergence of Yahoo boys sub-Culture in Nigeria, *International Journal of Cyber Criminology*

Ajeet, S. P. (2014). Cyber Crime: Challenges and its Classification. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Vol. 3 (6)*. Available: www.ijettcs.org.

Aliyu, M., Tambuwal, A. B., Namahe, Y. U. (2020). Investigating Factors and Extenuation Strategies for Mobile Phone Use While Driving in Nigeria. *Caliphate Journal of Science & Technology (CaJoST), Vol. 2(2).*
Aneke, S. O., Nweke, E. O., Udanor, C. N., Ogbodo, I. A., Ezugwu, A. O., Uguwishiwu, C. H., & Ezema, M. E. (2020). Towards Determining Cybercrime Technology Evolution in Nigeria. International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS) Vol. IX, Issue IV, April 2020 | ISSN 2278-2540
Augustine C. Odinma, MIEEE (2010): Cybercrime & Cert: Issues & Probable Policies for Nigeria, DBI Presentation, Nov 1-2.
Awhefeada, U. V., & Bernice, O. O. (2020). Appraising the Laws Governing the Control of Cybercrime in Nigeria. *Journal of Law and Criminal Justice*, *8*(1), 30-49.
Ezeanokwasa, J. O. (2019). Child Pornography under the Cybercrimes Act 2015 of Nigeria: The Law its challenges. *African Journal of Criminal Law and Jurisprudence*, *4*.
Frank, I. and Odunayo, E. (2013). Approach to Cyber Security Issues in Nigeria: Challenges and Solution. International *Journal of Cognitive Research in science, engineering, & education (IJCRSEE), 1* (1).
Geoff, H., Anthony, P., Gopalakrishnan, S. and Manav, M. (2005). Trends in Spam Products and Methods. Conference on e-mail and Antispam. Available online at www.ceas.org
Kshetri, N. (2019). Cybercrime and cybersecurity in Africa. Journal of Global Information Technology Managament, 22:2, 77-81, https://doi.org/10.1080/1097198X.2019.1603527
Longe O. B & Longe F. A (2005): The Nigerian Web Content: Combating the Pornographic Malaise Using Content Filters. *Journal of Information Technology Impact*, Vol. 5, No. 2, pp. 5964.
Longe, O, Omoruyi, I & Longe, F (2005): Implications of the Nigeria Copyright Law for Software Protection. *The Nigerian Academic Forum Multidisciplinary Journal.* Vol. 5, No. 1. pp 7-10.
Majesty, H., Cyber Crime Strategy, S.o.S.f.t.H. Department, Editor. 2010, The Stationery Office Limited: UK. p. 42.
Makeri, Y. A. (2017). Cyber Security Issues in Nigeria and Challenges. *International Journal Advanced Research in Computer Science and Software Engineering*, *7*(4).
Olumide, O. O. and Victor, F. B. (2010): E-Crime in Nigeria: Trends, Tricks, and Treatment. *The Pacific Journal of Science and Technology*, Vol. 11 (1), May 2010 (spring).
Omodunbi, B. A., Odiase, P. O., Olaniyan, O. M., & Esan, A. O. (2016). Cybercrimes in Nigeria: Analysis, detection and prevention. *Journal of Engineering and Technology*, *1*(1), 37-42. http://engineering.fuoye.edu.ng/journal/index.php/engineer/article/
Osho, O., & Adepoju, S. A. (2016). Cybercafés in Nigeria: Curse to the Internet. International Conference on Information and Communication Technology and Its Applications *ICTA 2016*, 117-123.
Roshan, N., What is cyber Crime. Asian School of Cyber Law, 2008: Access at – http://www.http://www.asclonline.com/index.php?titl e=Rohas_Nagpal,
Sodiq, K. A. (2012). Assessment of the Management of ICT Infrastructure of Selected Cybercafes in Lagos State. *Journal of Educational and Social Research*, *2*(9), 181-181.

Proposed Method on Phishing Email Classification Using Behavior Features

Ahmad Fadhil Naswir[a]*, Lailatul Qadri Zakaria[b], Saidah Saad[c]

[a,b,c] Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia
* Email: afadhilen@gmail.com

Abstract

Phishing email are also known as cyber-attack cannot be separated from the existence of the sender, the attacker of deceptive phishing will create an email based on observations and different ways of writing. Due to the characteristic of spam and phishing email constantly changed and updated, some features are needed to be modify to get better result. Generally, features used on the phishing email classification are the structure of the email itself namely email header, email body, and URL. With the large number of attackers who make email in the same scope on behalf of a trusted company, we can observe and analyze the differences in human/attacker aspect on the email content in terms of writing, word choice, and language style or in terms of stylometric features. There is still uncertainty of those features combined with email features on deceptive phishing email classification and detection. In other word, there are still human behavior features that can be observed more on deceptive phishing email classification, e.g. word choices, grammar, emotion of context, etc. that can be combined and implemented with email behavior features (header, body, and URL).

Keywords: Phishing Email; Phishing; Features; Features Selection; Behavior Features; Email Features; Stylistic; Stylometric; Email Classification

1. Introduction

Phishing email attacks are email based attacks which sent from someone or group of people for the purpose of fool the victim. It frequently will provide a luring message to trap the victim into entering at a fake website which look like a decent website. The website which already created by the attacker before doing the phishing attack will be provide some sensitive input, like username, password, credit card number, or any confidential data (Dakpa & Augustine, 2017). This type of email will be sent to predetermined target or massively to anyone depends on how the phishing will be done. Victims of the attacks will be tricked to input their confidential data to the fake website because the email tend to spoofed as an email from trusted companies such as Amazon, E-bay, or Google. The FBI has suggested that the impact of phishing attacks could be costing US businesses somewhere around $5 billion a year (Bagui et al., 2019).

Due to the characteristic of spam and phishing email constantly changed and updated, some features are needed to be modify to get better result. Generally, features used on the phishing email classification are the structure of the email itself namely email header, email body, and URL. Those features will be selected and extracted for improving the classifier performance (El Aassal et al., 2020). Phishing email cannot be separated from the existence of the sender, the attacker of deceptive phishing will create an email based on observations and different ways of writing. They utilize a strategy to create urgent atmosphere that convince the victim to react, for example, account alert or promising reward (Imaduddin et al., 2019). The attacker behavior aspects on writing part make as sure as possible to trap the victims to follow the flow of the email made by the attacker. Due to the large number of attackers who make email in the same scope on behalf of a trusted company, we can observe and analyze the differences in email content in terms of writing, word choice, and language style or in terms of stylometric features (Kumar et al., 2018).

2. Literature Review

2.1 Literature Review on Phishing Email Classification Features

There are several types of email phishing attacks which are: Deceptive Phishing, Spear Phishing, and Whaling (Birlea, M.C., 2020). In deceptive phishing attacks, the common attack occurred mostly by using email. The differences in email content in terms of writing, word choice, and language style or in terms of stylometric features. On authorship identification, stylometric features are being used for identify the authorship of an object mostly constructed by text (Kumar et al., 2018). There is still uncertainty of stylometric features combined with email features on deceptive phishing email classification and detection. In other word, there are still human behavior features that can be observed more on deceptive phishing email classification, e.g. word choices, grammar, emotion of context, etc. that can be combined with email behavior features (header, body, and URL) (Xiujuan et al., 2019).

In the current research on classification area, features selection and extraction are the most important aspect on how good the result of the classification will be. In text classification, there are several types of features representations used on the research one of which is on linguistic aspect that covers the writing, grammar, word choices, and tones on the text part of the email. Technique that commonly used on email classification is using machine learning or deep learning approach. For machine learning technique, features engineering part are extracted manually from the dataset into the classifier, as for deep learning technique, embedding is one of the methods to convert the features into a vector space model. As for the representation of the document, the behavior features that has selected will be transformed into an embedding or vector form as one of deep learning requirement to process the selected features (Gomez Adorno et al., 2018). In general, the combination of behavior features that selected will become the main feature for determining an email is categorized as phishing or not phishing either in machine learning or deep learning approach.

2.2 Literature Review on Human Behavior Features

There are some behavior features that used by several previous research on the area of classification. Commonly, the human behavior features that has been used are generally included in the stylometric class. Stylometry is an analysis of features that can be quantified such as sentence length, vocabularies, and frequencies. Therefore, any text related that can be measured is classified as stylometric features (Gomez Adorno et al., 2018).

Stylometric features are divided into two categories: low-level and high-level feature. Low-level features cover the number of words, characters, n-grams, etc.) and high-level feature is linguistic features (rhythmic, grammatical, tones). Each of the categories have sub categories which are word-based and character-based depend on the context provided (Lagutina et al., 2020). Typically, set of stylometric features used are divided into five categories (Sharon Belvisi et al., 2020):

Lexical: Set of characters and words (e.g. character count, word count, vocabulary richness), **Structural**: The way writer organizes the element in text. (e.g. lines count, paragraph count, etc), **Content-specific**: Frequency of particular/specific keyword in text, **Syntactic**: Syntax of the text. (e.g. punctuation, function words), and **Idiosyncratic**: Capture unique element of author. (e.g. misspelt word).

3. Proposed Method

From the explanation on section 2 and 3 above, it can be seen that there has been no significant impact in identifying deceptive phishing emails using stylometric features, which are generally used for authorship identification. The research methodology in this study is based on experimental research and focused on determining the best behavior features on detecting phishing email. By processing the dataset and using measurable and observable features, it is continued by conducting various experiments to obtain satisfying

results and be able to answer research questions and meet all the criteria of research objectives. In the design of this study, the research methods in conducting the testing process are as follows.

The following is a brief description of the phases of this research: The first phase is about Dataset Collection & Preparation, second phase is preprocessing, third phase is feature selection and extraction, next phase is feature embedding, fifth phase is deep learning algorithm implementation, and the last phase is enhanced classification model. On the third phase, email body text and the header will be analyzed and processed according to the stylometric features category. Each category (lexical, structural, etc.) will be converted into vector for the preparation on the next phase. The values of the features are depending on what are the content of the selected features. For example, on the lexical features will be extracted in the form of characters or words (e.g. number of words, number of characters, number of capital letters, etc.) (Kumar et al., 2018).

Based on the context of the features, there are some features that needed to be converted into numerical values that can be accepted for machine learning/deep learning process. One of them is by creating dictionary and applying one-hot encoding which convert text into binary number. Besides that, you can also use other methods by using n-gram. By counting the frequencies of each n-grams, the value can be used to represent the document as a vector. N-grams also can work for finding misspelling, language difference, and presence of other symbols in the text. Therefore, the extracted value from each of the selected feature will combined into one feature vector/embedding that can be inputted as the training data for the deep learning approach. Each of the selected feature will have its own value for determining the outcome of the classification process. By conducting several experiments, it is hoped that the behavior features with the best results can be identified.

4. Future Direction

Based on related work, there are several drawbacks from previous research regarding on technique, feature extraction and selection. There are researches used behavior features as the main features for the phishing email classification. Xiujuan et al. (2019) used 3 human behavior features namely stylometric, gender, and personality, Kumar et al. (2020) used linguistic features and URL features for the detection of phishing email. Based on the research above, there are more behavior features that can be observed more for improving the phishing email detection. For example, the grammar and typo from the email content could also be categorized as the human behavior feature.

On the past few years, there are several researches used deep learning method for email classification and the result is better than machine learning. Fang et al. (2019) develop a phishing email classification models based on RCNN and shows the performance result with the accuracy of 99.84% by using unbalanced dataset from Nazario and Enron email corpus. As for the machine learning approach, Kumar et al. (2018) used k-NN got highest accuracy of 95.48% and Kumar et al. (2020) used RF as the classifier, the experiment result is 97.75% of accuracy. From the result above and by excluding the features used on the research, the deep learning approach has the highest accuracy for phishing email classification. As the area of phishing email classification, features selection has become important part on determining the email is phishing or not phishing. Combination of several features can be implemented to produce a good result and accuracy. By understanding that feature selection is very influential in the continuity of experiments in email classification is a fairly advanced challenge.

5. Conclusion

The main scope of this paper covers the proposed method for classification and detection on phishing email with combination of behavior features used for the main parameter for classification. The novelty of this research is identifying the best combination of human and email behavior features for phishing email classification. Combined the selected features into feature embeddings for the data representation on phishing email classification. Finally, improved phishing email classification model with selected behavior features.

Acknowledgements

This research was supported by Universiti Kebangsaan Malaysia under research code [GGP-2020-041 UKM].

References

Bagui, S., Nandi, D., Bagui, S., & White, R. J. (2019). Classifying Phishing Email Using Machine Learning and Deep Learning. 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). Published. https://doi.org/10.1109/cybersecpods.2019.8885143

Birlea, M.C. (2020). Phishing Attacks: Detection and Prevention. ArXiv, abs/2004.01556.

Dakpa, T., & Augustine, P. (2017). Study of Phishing Attacks and Preventions. International Journal of Computer Applications, 163(2), 5–8. https://doi.org/10.5120/ijca2017913461

Das, S., Kim, A., Tingle, Z., & Nippert-Eng, C. (2019). All About Phishing: Exploring User Research through a Systematic Literature Review. ArXiv, abs/1908.05897.

El Aassal, A., Baki, S., Das, A., & Verma, R. M. (2020). An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. IEEE Access, 8, 22170–22192. https://doi.org/10.1109/access.2020.2969780

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. IEEE Access, 7, 56329–56340. https://doi.org/10.1109/access.2019.2913705

Gomez Adorno, H. M., Rios, G., Posadas Durán, J. P., Sidorov, G., & Sierra, G. (2018). Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. Computación y Sistemas, 22(1). https://doi.org/10.13053/cys-22-1-2882

Imaduddin, H., Widyawan, & Fauziati, S. (2019). Word Embedding Comparison for Indonesian Language Sentiment Analysis. 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT). Published. https://doi.org/10.1109/icaiit.2019.8834536

Kumar, A., Chatterjee, J., & Díaz, V.G. (2020). A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. International Journal of Electrical and Computer Engineering, 10, 486-493.

Kumar, S., Faizan, A., Viinikainen, A., & Hamalainen, T. (2018). MLSPD - Machine Learning Based Spam and Phishing Detection. Computational Data and Social Networks, 510–522. https://doi.org/10.1007/978-3 030-04648-4_43

Lagutina, K., Lagutina, N., Boychuk, E., & Paramonov, I. (2020). The Influence of Different Stylometric Features on the Classification of Prose by Centuries. 2020 27th Conference of Open Innovations Association (FRUCT). Published. https://doi.org/10.23919/fruct49677.2020.9211036

Sharon Belvisi, N. M., Muhammad, N., & Alonso-Fernandez, F. (2020). Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features. 2020 8th International Workshop on Biometrics and Forensics (IWBF). Published. https://doi.org/10.1109/iwbf49977.2020.9107953

Xiujuan, W., Chenxi, Z., Kang-feng, Z., Haoyang, T., & Yuanrui, T. (2019). Detecting Spear-phishing Emails Based on Authentication. 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS).

Analisa Corak Kunjungan Pesakit Luar Di Klinik Kerajaan Terpilih Negeri Selangor

Analysis of Outpatient Visit Pattern for Selected Government Health Clinics in Selangor

Suhaila Zainudin[a*], Dzulhusni bin Anjang Ab. Rahman[b]

[a]Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor
[b]Bahagian Pengurusan Maklumat, Kementerian Pendidikan Malaysia, Aras 3 & 4 Blok E11, ompleks E, Pusat Pentadbiran Kerajaan Persekutuan, 62604 Putrajaya
*Email : suhaila.zainudin@ukm.edu.my

Abstrak

Anggaran waktu menunggu sukar diramal kerana tempoh menunggu berkenaan bergantung kepada jenis rawatan kepada pesakit, prosedur yang perlu dilalui dan pendekatan yang digunakan oleh doktor bertugas dalam mendiagnosis serta seterusnya merawat pesakitt. Secara umum, klinik awam sentiasa penuh dan sesak pada waktu puncak. Namun berdasarkan set data TPC, terdapat corak berbeza waktu puncak antara klinik-klinik kerajaan yang berbeza dalam daerah yang sama. Corak-corak tersembunyai ini perlu dikenalpasti terlebih dahulu bagi membolehkan peramalan waktu menunggu dan mencadangkan kepada pesakit klinik mana yang terbaik untuk dikunjungi pada waktu tertentu bagi mengelakkan tempoh menunggu yang lama.

Katakunci: waktu menunggu; analisisi deskriptif; penerokaan pengetahuan

Abstract

The estimated waiting time is difficult to predict as the waiting period depends on the type of treatment to the patient, the procedures to be performed and the approach used by the doctor to diagnose and subsequently treat the patient. In general, public clinics are always full and crowded at peak times. However, based on the TPC data set, there are different peak hours patterns between different government clinics in the same district. These hidden patterns need to be identified in advance to enable the forecasting of waiting times and suggest to patients which clinics are best to visit at certain times to avoid long waiting periods.

Keywords: waiting time; descriptive analysis; knowledge discovery.

1. Pengenalan

Tempoh menunggu yang panjang adalah isu yang lazim untuk pengguna perkhidmatan kesihatan di klinik kesihatan awam (Ahmad et al. 2017). Penyelidikan bagi mendapatkan pendekatan menyeluruh yang efektif dan efisien bagi menyelesaikan isu ini giat dilaksanakan (Morley et al. 2018). Masa rawatan berbeza mengikut negara dan ditentukan oleh ciri-ciri pesakit dan doktor (Ahmad et al. 2017).

Sebagai salah satu usaha mempertingkatkan tahap perkhidmatan kesihatan di klinik kesihatan atau KK, Kementerian Kesihatan Malaysia telah melaksanakan pelbagai inisitif berbentuk ICT dan bukan ICT. Sistem Teleprimary Care (TPC) merupakan salah satu inisaitif berbentuk ICT dengan tujuan melancarkan proses menunggu di klinik kesihatan awam. Sistem TPC ini merupakan sistem rekod perubatan elektronik yang

merekodkan aktiviti pesakit luar dari daftar masuk sehingga selesai mendapatkan perkhidmatan (Ho 2013). Set data TPC untuk beberapa klinik kesihatan di Selangor telah diperolehi sebagai input kajian ini.

1.1. Latarbelakang

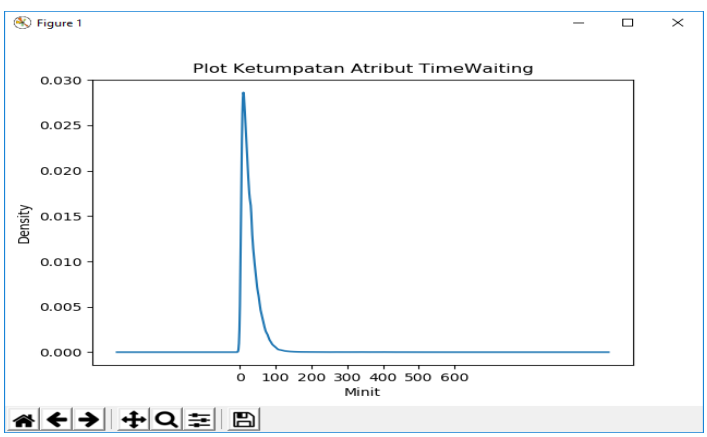
Ahmad et al. (2017) menyatakan bahawa menunggu adalah fenomena lazim di ruang menunggu doktor. Tempoh menunggu yang lama adalah pembaziran dari segi sumber manusia, wang, dan masa serta meningkatkan kekecewaan pesakit (Chen et al. 2016). Meramalkan tempoh menunggu pesakit luar di hospital atau klinik adalah aktiviti yang kompleks. Tempoh menunggu dipengaruhi oleh proses yang dilalui oleh setiap pesakit yang berbeza antara satu sama lain. Tempoh menunggu ini juga bergantung kepada proses yang dilalui oleh pesakit yang hadir lebih awal atau pesakit dengan keutamaan tertentu yang dibenarkan untuk memotong barisan menunggu. Terdapat juga faktor luaran yang tidak dapat dikawal seperti pendekatan setiap pegawai perubatan yang mendiagnos pesakit (Ahmad et al. 2017), keupayaan pesakit menerangkan simptom penyakitnya kepada pegawai perubatan, masalah berkaitan dengan bahasa pertuturan dan kepelbagaian ujian saringan yang perlu dilalui oleh pesakit (Chen et al. 2016).

2. Analisis Deskriptif Untuk Set Data Daerah Klang

Analisis deskriptif adalah merangkumi pengumpulan data, penyusunan data, ringkasan data dan visualisasi data. Ianya bertujuan untuk menerangkan ciri-ciri sesebuah set data. Contoh analisa deskriptif adalah mencari nilai mean, median, jumlah, varians, nilai minimum, nilai maksimum, skewness dan sebagainya. Set data itu kemudian diwakilkan dalam bentuk graf dan carta bagi meningkatkan kefahaman (Loeb et al. 2017). Bagi kajian ini, penerangan tentang analisis deskriptif untuk tiga klinik kesihatan di daerah Klang akan diterangkan. Klinik Kesihatan(KK) diterangkan untuk KK Anika, KK Bukit Kuda dan KK Pandamaran. Hasil analisa deskriptif set data daerah Klang dengan menggunakan fungsi Python describe ( ) seperti dalam Jadual 1.

Jadual 1. Keterangan set data daerah Klang

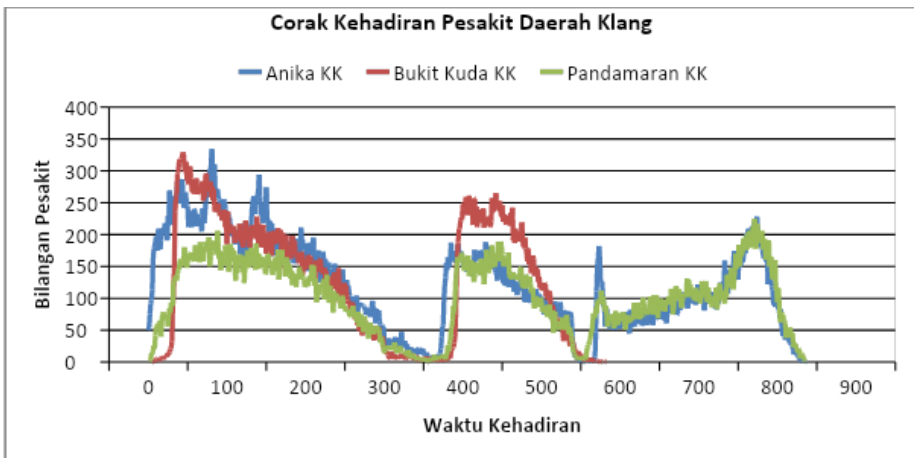
	DayOfMonth	Month	ArrivalHour	ConsultationLength	TimeWaiting
Jumlah	261,297	261,297	261,297	261,297	261,297
Purata	15.85	6.47	12.57	5.47	28.44
Sisihan Piawai	8.73	3.46	4.02	20.38	26.32
Min	1.00	1.00	7.00	-765	0
Maks	31.00	12.00	21.00	701.00	687.00



Rajah 1. Plot ketumpatan atribut TimeWaiting daerah Klang

2.1 Corak Ketibaan Pesakit Bagi Daerah Klang

Rajah 2 menunjukkan jumlah pesakit berdasarkan waktu ketibaan ke klinik-klinik kesihatan di daerah Klang. Secara umumnya corak ketibaan pesakit di ketiga-tiga klinik kesihatan berkenaan menunjukkan corak yang sama dari sudut masa puncak dan bilangan pesakit.

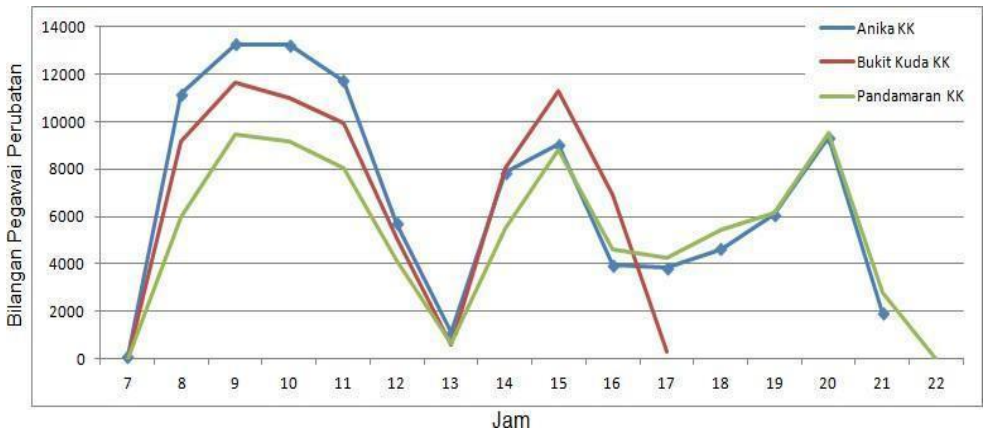


Rajah 2. Corak waktu ketibaan pesakit bagi klinik-klinik kesihatan di daerah Klang

Pesakit mula tiba di KK Anika seawal jam 7.30 pagi. Ini dapat dilihat dari arah aliran menaik dan kecuraman graf. Terdapat enam waktu puncak iaitu pada sekitar jam 8.00 pagi, 8.45 pagi, 9.50 pagi, 2.00 petang, 5.00 petang dan 10.00 malam. Puncak tertinggi ketibaan pesakit bagi klinik kesihatan ini ialah pada jam 8.45 pagi. Bagi Klinik Kesihatan Pandamaran pula pesakit juga mula tiba sekitar jam 7.30 pagi dengan arah aliran menaik tetapi tidak mendadak seperti KK Anika. Corak ketibaan pesakit bagi KK Bukit Kuda agak berbeza berbanding dengan kedua-dua klinik yang dihuraikan sebelum ini di mana klinik ini menerima pesakit yang paling ramai bagi antara pulul 2 hingga 4 petang.

Pesakit mula tiba selepas jam 7.30 pagi dan arah aliran mula menaik bermula jam 8.00 pagi dengan puncak tertinggi kehadiran pada jam 8.15 pagi. Titik puncak seterusnya pula adalah pada jam 8.40 pagi, 9.45 pagi, 2.00 petang dan. 2.50 petang. KK Bukit Kuda tidak dibuka pada waktu malam dimana ketibaan pesakit terakhir adalah sebelum jam 5.30 petang.

Walau bagaimanapun, jumlah pesakit yang mengunjungi KK Pandamaran adalah kurang berbanding dengan Klinik Kesihatan Anika dan KK Bukit Kuda. Waktu puncak bagi klinik ini ialah pada jam 9.00 pagi, 2.00 petang, 3.00 petang, 5.00 petang dan 10.00 malam. Rajah 3 menunjukkan bilangan pegawai perubatan bertugas dan jumlah konsultansi mengikut jam bagi tahun 2018. Jumlah konsultansi kepada pesakit luar mencapai nilai maksimum diantara jam 9.00 pagi sehingga jam 10.00 pagi dan mula menunjukkan corak menurun selepas itu. Pada sebelah petang pula, bilangan konsultasi maksimum oleh pegawai perubatan bertugas pada jam 3.00 petang dan pada sebelah malam pula pada jam 10.00 malam. Ini menunjukkan terdapat perbezaan jumlah pegawai perubatan yang bertugas mengikut jam. Perkara ini berkemungkinan disebabkan oleh perbezaan waktu bekerja yang bermula seawal jam 7.30 pagi sehingga jam 8.30 pagi. Pengurangan jumlah konsultansi oleh pegawai perubatan bertugas selepas jam 10.00 pagi tidak dapat dijelaskan.



Rajah 3 Jumlah pegawai perubatan mengikut jam

3. Kesimpulan

Hasil daripada analisa deskriptif dan visualisasi set data menunjukkan terdapat persamaan corak set data antara klinik-klinik kesihatan walau pun klinik berkenaan berada dalam daerah yang sama. Corak kedatangan pesakit luar juga berbeza bagi setiap klinik kesihatan berdasarkan waktu ketibaan dan waktu puncak. Corak ini boleh dijadikan panduan dalam penetapan waktu bekerja dan jadual pergerakan pegawai perubatan di mana penjadualan semula pesakit dan pegawai perubatan boleh dilaksana untuk mengoptimum sumber yang sedia ada.

Penghargaan

Kajian ini disokong oleh Geran Universiti Penyelidikan UKM (GUP-2020-089).

Rujukan

Ahmad, B. A., Khairatul, K., & Farnaza, A. (2017). An assessment of patient waiting and consultation time in a primary healthcare clinic. Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia, 12(1), 14.

Chen, J., Li, K., Tang, Z., Bilal, K. & Li, K. (2016). A parallel patient treatment time prediction algorithm and its applications in hospital queuing-recommendation in a big data environment. IEEE Access. doi:10.1109/ACCESS.2016.2558199

Ho, Ai Chia. (2013). Acceptance and utilization of teleprimary care in Sarawak. PhD thesis, Universiti Malaysia Sarawak. Available from https://ir.unimas.my/id/eprint/15439/.

Loeb, S., Dynarski, S., Mcfarland, D., Morris, P., Reardon, S., Reber, S. & Shodganga. (2017). Chapter 7 Descriptive Analysis 7.0 Chapter Overview (March): 39. doi:10.1016/B978-1-4377-0651-2.10007-4.

Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS One. 2018;13(8):e0203316. Published 2018 Aug 30. doi:10.1371/journal.pone.0203316.

House Price Prediction in Selangor Using Machine Learning Algorithms

Azwanis Abdosamad [a*], Nor Samsiah Sani[b]

[a,b]Faculty of Information Science and Technology,Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor
*Email: P95394@siswa.ukm.edu.my

Abstract

The increase of housing price every year is very worrying, especially for buyers who are in urban areas. Selangor has an area with a high population density and high house prices prompted for this study to be conducted. Particularly, it is very helpful for house buyer/seller or a real estate agent broker get insight in making wise decision considering the housing price prediction. The purpose of this study is to find out and identify the important features that influence housing price in Selangor. The housing dataset is obtained from National Property Information Centre (NAPIC) which has a total of 64982 data and 23 attributes. The dataset contains data of residency sector in Selangor from 2015 through 2020. The house price attribute was selected as the dependent variable which is the target value in this study. After feature selection is made, several parameters were optimized to model the house price prediction. Three (3) algorithms are Random Forest (RF), Gradient Boost Decision Tree (GBDT) and k-Nearest Neighbors (k-NN) are developed by using machine learning techniques. Mean Squared Error (MSE) values of each algorithm is determined and compared to find the best algorithm in term of accuracy. From the findings, it is found that RF algorithm achieves the best performance model with MSE 0.00017549 value

*Keywords***:** Machine Learning; Random Forest; Gradient Boost Decision Tree; k-Nearest Neighbors; Mean Square Error

1. Introduction

A house is defined as a home that meets other basic needs (UN-Habitat, 2011). In an era of advanced technology, homes are not only the shelter for people, but also a long-term asset and investment. However, the increase in house prices which is increasingly worrying in Malaysia which causes the people in this country not afford to own their own house (Azima Abdul Manaf, 2019). There are many factors that causing a serious increase in house prices, among them are the demand, supply of house prices and pricing by developers. A good forecast model is needed to predict house prices. Thus, house price prediction models using different machine learning algorithms to produce high-accuracy forecast models.

Main objectives of this paper are to identify the important features that influence the price of a house in Selangor, to develop and make a comparison of three (3) models by using machine learning techniques and identifying the best house price model among of the three models developed. In this study, data set is obtained from National Property Information Centre (NAPIC). The data is a type of residential category which includes terraced and multi-floor houses that not exceed three floors. This data set were collected in Selangor from 2015 until mid-2020 has a total of 64982 data with 23 attributes.

2. Related Work

Winky K.O et al. (2020) has used RF, GBDT, SVM to determine the house price prediction in Hong Kong by using those algorithms. He concludes that RF and GBDT produce more accurate price estimates than SVM.

Sifei Lu et al. (2017) attempt to estimate house price in Ames by using Lasso and Gradient Boosting Decision Tree. House price forecast based on house characteristics and location which is find out that GBDT produces better model capabilities than Lasso. Study conducted by Nissan, Pow, Emil Janulewics & Liu (2015) prove that k-NN and Random Forest show excellent performance compared to linear regression. k-NN is the best model with the lowest square error rate. Their study is using the data set which is extracted from the centris.ca website.

3. Material and Method

Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and k-Nearest Neighbors (k-NN) were applied for regression and comparison between them for the most accuracy model. After the analysis, conclusion and the recommendations have been write up to provide the output

3.1 Data Pre-Processing

In this study, Scikit-Learn and Microsoft Excel are being used as a machine learning tools to perform pre-processing tasks. In a data cleaning process, there is data missing and not filled in certain attribute lines. The solution is taken to manually fill in Excel. Missing data is fill in with the median value. In this study, the missing data on attributes that require filling in the median values are atribut luas_lot, luas_lot bangunan and b_tingkat. There are also attributes that have noise data where the data is filled in with incorrect or unreasonable values. The attribute is b_tingkat. Data reduction is made so that the remaining values are correct. There are also attributes that have no value in some lines and have no relation to other attributes such as the keadaan_bgn attribute. The solution is also to delete the data from the data record. There are 44 records of noise and irrelevant data where these records are being deleted. The data in the data set also goes through the process of data transformation where converting category data to numeric. After going through the pre-process, only 9 attributes are left that are called as essential attributes. The attributes are as in Table 1 below.

Table 1 Essential Attributes with Data Type

Essential Attribute	Data Type
daerah1	nominal
jenis_pegangan	nominal
pro_type	nominal
b_tingkat	numeric
luas_lot_bgn	numeric
luas_lot	numeric
harga_b	numeric
keadaan_bgn	nominal
thn_perjanjian	numeric

4. Results and Discussion

Nine important attributes that have been selected are tested using the feature selection method and then through the correlation matrix method to determine the strength of the relationship between those attributes and the house price attribute. Figure 4 shows the heat map used in the correlation matrix experiment using numerical attributes. The heat map shows the correlation relationship between the attributes and the house price where the

relationship strength is shown in the form of numerical values. The relationship between attributes and house price attributes is shown in Table 2 below. The correlation value of this coefficient is between -0 and 1.

Table 2 Level of Attributes Relationship with Home Price Attribute

Attribute	Correlation Matrix	Level of Relationship
b_tingkat	0.34	Quite strong
luas_lot	0.16	Moderate
luas_lot_bgn	0.16	Moderate
thn_perjanjian	-0	No correlation

In this study after the experiments were conducted, the Random Forest (RF) algorithm gave the lowest MSE value of 0.00017549, followed by the Gradient Boosting Decision Tree (GBDT) algorithm and k-Nearest Neighbors (k-NN) with a value of 0.00020321 and 0.00022385 respectively. The result of each algorithm is shown in Table 3 below.

Table 3 Level of Attributes Relationship with Home Price Attribute

Prediction Model	MSE
RF	0.00017549
GBDT	0.00020321
k-NN	0.00022385

5. Conclusion

This study employs machine learning techniques to develop a price prediction model for house in Selangor. It uses a dataset of residential housing for a 5-year period from year 2015 until 2020. The regression model performances of the models are compared with one another and the accuracy of the prediction in this study assessed by checking the mean squared error score of the training model. The pre-processing method has been made before the test and the data is divided into two parts which is the training set and the test set. Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and k-Nearest Neighbors (k-NN) model which is a machine learning regression algorithm are used in this study. The correlation relationship between the dependent and independent variables was determined through the Pearson Coefficient. In this relationship is explained through a correlation matrix where the relationship is expressed in coefficient values. From the results of the experiments conducted, the correlation relationship for the numerical type variable is moderate with the house price dependent variable. Based on the experimental results, the Random Forest (RF) model is the best predictive model and has high accuracy compared to the Gradient Boosting Decision Tree (GBDT) and k-Nearest Neighbors (k-NN) models.

Acknowledgements

I would like to acknowledge with gratitude to my supervisor Dr Nor Samsiah for her guidance in the process of this work. Special thanks to National Property Information Center (NAPIC) for the data set used in the experiment.

References

Abdul Razak, H., Azuraliza, A. B. & Mohd Zakree, A. N. (2018). Sains Data Penerokaan Pengetahuan dari Data Raya, hlm. Ed. Ke-1. Malaysia: Penerbit Universiti Kebangsaan Malaysia (UKM).

Azima Abdul Manaf & Goh le Zheng. (2019). Faktor-faktor Penentu Harga Rumah dari Perspektif Pemaju Perumahan. Malaysian Journal of Society and Space 15 issue 4 (246-260).

Pow, Nissan, Emil Janulewicz & L. Liu. (2015). Applied Machine Learning Project 4 Prediction of Real Estate Property Prices in Montréal.

Sifei Lu, Xulei Yang & Zengxiang Li. (2017). A Hybrid Regression Techniques for House Price Prediction. *ResearchGate*.

United Nations Human Settlements Programme (UN-Habitat). (2011). A Practical Guide for Conducting: Housing Profiles Supporting Evidence-Based Housing Policy and Reform. ISBN Number: 978-92-1-132028-2.

Winky,K.O & Bo-Sin Tang. (2020). Predicting Property Price with Machine Learning Algorithms. *Journal of Property Research*.

Analyzing Twitter Reviews on Halal Food using Sentiment Analysis

Alya Nur Adlina Ahmad Nazri[a], Siti Nur Kamaliah Kamarudin[b]

[a,b] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
* Email: alyaadlinazri@gmail.com; snkamaliah@uitm.edu.my

Abstract

Malaysia have among the highest number of social media users, and one of the popular topics discussed online is the halal status of trendy restaurants or popular food. Since most users nowadays use social media like Twitter to express their opinions, it is also convenient for them to search information about the halal status of their place or food of interest. However, it can be time consuming for users to confirm the halal status of some places or food in Malaysia as some places may serve halal food but they do not bother to get the Malaysian halal certification. Therefore, this research explores the process of how to evaluate the sentiments on halal food or restaurants from Twitter tweets and to identify the halal, non-halal or pork-free status of the food or restaurant. A sentiment analysis was performed using a Lexicon-based approach to predict and label the tweets. Subsequently, Machine Learning classifiers such as Support Vector Machine (SVM), Naïve Bayes, and Deep Learning were applied to compare the sentiment analysis performance. The total number of tweets used in this research was 71,333 tweets where as a result, Deep Learning method achieved the highest accuracy with 73.18% using SentiWordNet on N-grams range of [1,3]. From these findings, Deep Learning outperformed other classifiers as it scores the highest for both accuracy and f1-score metrics.

Keywords: halal food; Twitter reviews; sentiment analysis; machine learning; lexicon-based approach

1. Introduction

The halal industry are synonymous with areas like tourism, cosmetics and of course, food. Other industries such as pharmaceuticals and healthcare goods have also started to apply the concept of halal (Sulaiman et al. 2018) due to the increase of demand for halal items. Twitter has been one of the most popular social media platforms to be used for sentiment analysis (Tyagi & Tripathi 2019). It's functionalities is similar to a huge forum where users can contribute their ideas, thoughts, and opinions making it an ideal platform to analyze different opinions on various areas (Sarlan et al. 2015). In 2019, around 2.5 million Malaysians were active users on Twitter (*Malaysia: number of Twitter users 2014-2019* 2019) where the users posted, tweeted and shared many opinions on various issues. One of the commonly debated issues on Twitter is on the *halal-ness* of a certain food or restaurant, especially if the food or restaurant suddenly became popular in the country (Feizollah et al. 2019).

The huge number of Malaysian Twitter users expressing their personal opinions via twitter benefits those who would like to search or read reviews on certain food or restaurant. However, it can also be time-consuming to go through all of the reviews. Additionally, some tweets may use short forms or unfamiliar terms which makes it even more time-consuming for users to sift through the tweets to find the halal status for a particular food or restaurant. Therefore, for this research the author chose to investigate and classify tweets on halal food reviews taken from Twitter in order to help users search for the halal status of a particular food or restaurant. Subsequently, this research attempted to identify some of the keywords used to determine the halal status of the food in the tweets by users. This research also make use of several machine learning approaches to classify the tweets and finally, results will be displayed through a dashboard which was incorporated into *Halalopedia*, a web-based system created using the results from this research.

2. Literature Review

1.1 Halal Food Certification in Malaysia

Nowadays, Muslims have a variety options of halal products, food and beverages and services offered. However, Nurrachmi (2018) has reported that halal food suppliers mostly came from non-Muslim countries like New Zealand, Australia, France, and Canada. This has shown that countries with lesser Muslim populations are well aware of the halal sources. In Malaysia, halal goods are recognized by searching for a halal logo issued by JAKIM or any other halal certified organization. Besides, a considerable amount of literature regarding halal food have been published. In a study, it was found that roughly 70% of Muslims all around the world adhere to at least some of the halal food restrictions (Ahmad et al., 2018).

1.2 Sentiment Analysis

Sentiment is known as the opinions expressed by individuals that contain feelings, attitudes, and thoughts. Sentiment analysis analyses textual context using natural language processing and classifies it as positive, negative, or neutral (Hassan 2019). It was broadly applied to analyze how people feel about something based on their sentiments. According to Chen & Zhang (2018), sentiment analysis generally uses natural language processing (NLP), text interpretation, machine learning, computational linguistics, and other approaches to interpret, process and trigger emotionally colored messages. The two most widely used methods to conduct sentiment analysis is by using machine learning approach or lexicon-based approaches (Sarlan et al. 2015).

- *Machine Learning Approach:* According to Hasan et al. (2018), machine learning approach was essentially intended to identify textual content by implementing algorithms like naïve bayes and support vector machine (SVM). Naïve bayes, deep learning and support vector machine are examples of supervised machine learning algorithms while k-means is unsupervised algorithms. The goal of supervised learning is to predicts the final outcome variable using the predictor variable. Moreover, supervised learning aims at automating time-consuming, or costly manual tasks (Mittal & Patidar 2019).
- *Lexicon-based approach:* Lexicon-based approaches are part of unsupervised learning algorithms. Using this approach, the positive and negative words in dictionary will match the words in the tweet. These techniques, however, depends entirely on lexical resources that are concerned with mapping words to a score of categorical, or numerical sentiments. Additionally, lexicon-based approaches require no training data, and depends solely on dictionary. The sentiment lexicon comprises an index of words and contains the polarity details of the related terms, whether positive or negative. However, the limitation of lexicon dictionary was, not all words in the sentiment can be assigned with a value (Sarlan et al. 2015).

3. Methodology

3.1 Data Collection and Pre-processing

Tweets were collected by scrapping from Twitter using TWINT module, using the Twitter search function related to halal food and restaurant from recent years. This project did not use Twitter API for data collection as even though it is the most conventional method the extract data from Twitter, it has many limitations like limited time span and limited access to Twitter server. Total data scraped using TWINT were approximately 72,000. Tweets were also collected using the keywords identified. The dataset consists of details such as Tweet Id, time and date of tweet, and location of the tweet. Several data pre-processing activities were conducted to achieve the cleaned data set, such as data transformation, filtering, tokenization, normalization, and application of N-gram. Duplication of tweets were also performed on dataset using Rapid Miner software. And finally, the

dataset was labeled to positive, neutral and negative sentiments based on the polarity score through Rapid Miner software as well.

3.2 Modelling and System Development

Three machine learning classifiers were applied to compare the accuracy and the performance results using RapidMiner software. The classifiers chosen were SVM, Deep Learning, and Naïve Bayes where the dataset were split to a 90:10 ratio. Performance metrics taken into account for this research were accuracy and f1-score. Subsequently, the results were published through a dashboard created using Power BI software and through a web-based system named *Halalopedia* created using PHP and HTML language.

4. Results and Discussion

4.1 Machine Learning Classifiers Performance

In our results, SVM and Naïve Bayes works better with Vader approach and [1,3] N-grams range with 68.44% and 67.03% accuracy respectively. Meanwhile, deep learning works best with SentiWordNet approach on [1,2] N-grams range. Besides, roughly, [1,2] range gives out the lowest accuracy except for SVM + SentiWordNet and NB + SentiWordNet. By comparing the accuracy for each model, deep learning achieved the highest accuracy with 73.18% using [1,2] N-grams range and SentiWordNet. Contradictory, SVM and naïve bayes achieved highest f1-score when it is used with SentiWordNet approach with 56.28% and 57% respectively whereas deep learning scored highest using Vader approach with 58.16%. However, all the highest f1-score for each classifier is achieved when performed on [2,3] N-grams range. Since deep learning achieved highest for both accuracy and f1-score metrics, it can be concluded that deep learning classifier is the best model for this project.

To summarize, deep learning is found to produce better accuracy compared to SVM and naïve bayes with highest accuracy of 73.18%. Besides, deep learning results are also much better when used with SentiWordNet approach. The only downside of deep learning is that the processing time took the longest to finish which approximately around 20 minutes while other classifiers is between 10-15 minutes. Additionally, it is also proven that deep learning can overcome short texts dataset problem. However, the dataset used in this project are imbalanced for each class label from the confusion matrix. Therefore, for this case, accuracy might not be the best performance measure. The accuracy on imbalanced data have mislead the performance of the sentiment analysis. Hence, f1-score is a good metric when the data is imbalanced as it considers the precision and recall value of the data. Considering the f1-score, the highest score is considered to be the best model for imbalanced data as it can predict better on multiclass classification.

4.2 Results discussion

Based on the experiments done, it can be seen how pre-processing phase is very important and need to be done thoroughly. This is because, in Twitter there are many slang, dialect and short form words are used. Nevertheless, there are still some slang and Malay words that cannot be detected during data pre-processing, hence contributing to one of the reasons on why this project does not get high accuracy above 80%. Therefore, variety range of N-grams are used in this project to improve the sentiment analysis performance.

5. Conclusion and Future Works

In conclusion, this project managed to accomplish all the research objectives stated earlier. The result of sentiment analysis is able to analyze and determine the tweets on food reviews into halal, non-halal and pork-

free food. The dashboard is successfully developed to display the sentiment analysis results. In addition, the search page in web system also able to help user to find the information on the halal status of a food or restaurant. Overall, this project is still lacking in many ways and is open for more improvement in the future.

Acknowledgements

The authors would like to express their gratitude towards Faculty of Computer and Mathematical Sciences for the research opportunity.

References

Ahmad, A. N., Ungku Zainal Abidin, U. F., Othman, M., & Abdul Rahman, R. (2018). Overview of the halal food control system in Malaysia. *Food Control*, *90*, 352–363. https://doi.org/10.1016/j.foodcont.2018.02.035.

Chen, Y., & Zhang, Z. (2018). Research on text sentiment analysis based on CNNs and SVM. *Proceedings of the 13th IEEE Conference on Industrial Electronics and Applications, ICIEA 2018*, 2731–2734. https://doi.org/10.1109/ICIEA.2018.8398173.

Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., & Hazim, M. (2019). Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms. *IEEE Access*, *7*, 83354–83362. https://doi.org/10.1109/ACCESS.2019.2923275.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011.

Hassan, F. (2019). *Sentiment Analysis: Analyzing Online Food Reviews for Best Dishes*. *2018* (October 2018).

Malaysia: number of Twitter users 2014-2019. (2019). Statista Research Department. https://www.statista.com/statistics/490591/twitter-users-malaysia/

Mittal, A., & Patidar, S. (2019). Sentiment analysis on twitter data: A survey. *ACM International Conference Proceeding Series*, 91–95. https://doi.org/10.1145/3348445.3348466

Nurrachmi, R. (2018). The Global Development of Halal Food Industry: A Survey. *Tazkia Islamic Finance and Business Review*, *11*(1), 41–56. https://doi.org/10.30993/tifbr.v11i1.113

Sarlan, A., Nadam, C., & Basri, S. (2015). Twitter sentiment analysis. *Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014*, *November*, 212–216. https://doi.org/10.1109/ICIMU.2014.7066632

Sulaiman, M. Z. M., Noordin, N., Noor, N. L. M., Suhaimi, A. I. H., & Isa, W. A. R. W. M. (2018). Halal inspection process at federal and state level: A case study of Halal Certification system in Malaysia. *2017 IEEE Conference on Open Systems, ICOS 2017*, *2018-Janua*, 65–70. https://doi.org/10.1109/ICOS.2017.8280276

Tyagi, P., & Tripathi, R. C. (2019). A Review Towards the Sentiment Analysis Techniques for the Analysis of Twitter Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3368718

Digital Market Governance and Challenges on Competition Law in Asia: Malaysia, India, and Indonesia

Angayar Kanni Ramaiah [a*], Anupam Sanghi [b], Ningrum Natasya Sirait [c]

[a]University Teknologi MARA, Cawangan Pulau Pinang Malaysia, Bukit Mertajam, 35000 Malaysia
[b] Anupam Sanghi & Associates, India
[c]Universitas Sumatra Utara, Medan, Indonesia
* Email: kanni844@uitm.edu.my

Abstract

The magnitude of the consumer market in East Asia in particularly India, and ASEAN nations has evolved a robust digital revolution and competition, enhancing the efficiency of the nations digital market. This digital economy is also succumbed with multitude of unprecedented challenges, like the killer merger, algorithmic manipulations, and abuse of dominance. So, tension has grown on the governance and regulatory control on the usage of the technological tools and resulting impact on the consumer welfare. Although this is indeed a global phenomenon, the Asian regulators are further challenged because of their relatively inadequate, rather unprepared rules of the game in digital platforms, to regulate the unfathomable, digital features like, the big data (with technologically advanced tools) management, algorithmic usage, and related artificial intelligence (AI). These tech features have shifted beyond human sovereignty to machine efficiency (data gathering, processing and usage) bypassing all the conventional legal podium. The developing economy like ASEAN and India are heavily dependent on the Western technology and their Competition Law enforcement is comparatively still at infant stage with limited resources to tackle or legitimise all the digital activity. This paper firstly highlights the lack of jurisdictional control and challenges on Asian competition agencies, on digital mergers, algorithmic price setting mechanism and abuse of dominance with case study in Malaysia, Indonesia, and India. Secondly, discusses enforcement mechanism adopted in advanced jurisdiction like European Union and the United States before concluding the discussion with some recommendations for improvement.

Keywords: competition law; digital economy; big data; algorithm and artificial intelligence

1. Introduction

Digital marketing, with smart digital computing technologies, on platform-markets via internet and the World Wide Web (WWW) infused with various anti-competitive collusion (price-fixing or cartel), monopoly, and Mergers and Acquisition (M&A) using Artificial Intelligence (AI), as hub and spoke network and algorithm. The governance of this digital knowledge management and marketing is still an evolving experience for Asian regulators, who are rather slow and often challenged to properly address its atrocities. Thus, significant decline in 'competition' experienced in precipitating concentrated market or market dominance, with the rising market power of large firms that are slowing business dynamism (Foda & Patel, 2018) and transactional decisions of the consumers (when providing their personal data) (Ritter & Slyom, 2018). Therefore, potential digital firm's scale on platforms, its intangible capital, market concentration and knowledge monopoly must be moderated to ensure level playing field for competitors, newcomers as well as consumer welfare. Competition Law (CL) and policy that moderates market efficiency has a broad and deep role to moderate these fast-moving digital features from adversely effecting market competition (Peter & Singh,2019). This paper discusses the digital market challenges on Asian eco-system with respect to digital merger,

algorithmic price setting and abuse of dominance case analysis from Malaysia, Indonesia, and India and concludes with some recommendations for improvement.

2. Digital Knowledge Management, Technology, and Marketing Under Competition Law

2.1 Digital Economy and Competition Law

The task of legitimising digital corporations conducts under CL became critical issue of concern after various scandals and public upheavals surfaced with a number of important competition cases brought against the digital kings like Facebook, Google, Amazon, Apple, and Microsoft (FGAAM). FGAAM alleged for preserving and monopolising market power to undermine competition on merit (Fung,2020) and consumers welfare. The unprecedented scale and speed with which personal data is collected and used in the context of prediction algorithms, an omniscient, opaque machinery threatens to erode the very foundation of consumer privacy. And the ability of their digital monopolies to control much of our attention by dictating which content exposed to and to influence our behaviour, brings about the "economy of attention," users' where eyeballs become the main commodity traded, monetization for ads such as on YouTube or Facebook, ranging from a few cents to several dollars depending on the specificity of the target audience (Marz, et.al.,2018, April 10). Hence, there is robust debate regarding the extent to which CL should be applied on these digital features to regulate their markets.

CL enactments generally aimed to control or eliminate restrictive agreements or arrangements among enterprises, or mergers and acquisitions or abuse of dominant positions of market power, which limit access to markets or otherwise unduly restrain competition, adversely affecting domestic or international trade or economic development (UNCTAD,2010) to ultimately widen the consumer choice, reduce prices, and improve quality besides encouraging innovation (EU, 2012). CL policy emphasises businesses of the same kind to compete fairly with each other to maximize consumer welfare (Whish & Bailey, 2015). Generally, CL prohibits anti-competitive agreements horizontally (or cartel) or vertically (between economic players at different levels), abuse of market power by dominant firms or attempts of not yet dominant firms to monopolise markets and ant-competitive merger or acquisition (M&A). However, there is considerable divergence in each jurisdiction about the precise definition of prohibited agreement or dominance, the range of practices and conducts that should be condemned as anti-competitive, and remedies that should be imposed. And for many developing countries, CL itself relatively novel area and authorities have limited resources and experience in handling anti-competitive cases.

2.2 Challenges in Applying Competition law on Digital Economy

Digitalised activities challenge regulators and the existing legislative tools, to channel the CL infringement action on (non-human) algorithms or creators and/or users or the data abusers. Automation, AI, and algorithmic although efficient but seriously undermine human sovereignty, at every stage: in the manner data is collected and processed and stored, as well as way algorithms is used, the authority it is given to act, technical rules for the construction of algorithms and the supervision of automated execution. The sheer greed for data causes M&A, in the disguise of asset acquisition to increase concentration by (especially big companies) by buying out upstart rivals before even it becomes a competitive threat. Mergers may not only hinder price competition but also impede innovation and enable the dominant corporations to adopt harmful exclusionary practices such as refusals to deal with rivals, restrictive contracting, tacit collusion, and predatory pricing that ultimately squelch competition **(**News Release, 2021, February 4**)** and impair consumer welfare. Lack of expertise and

means (legal tools) to assess the harm has further reduced the pace of enforcement and success of the infringement charges to legitimise the fast-moving digital technology. The fault-finding procedures are faced with multitude problem in gathering intangible digitised based evidence, automated theorem (based on the Evidence Algorithm (EA) and System for Automated Deduction (SAD)). Hence, EA information and SAD access requires expertise and experts to gather the evidence for the legal proceedings. The finding unveiled this

3. Digital Economy in Asian Eco-system: Governance Issues and Challenges

3.1 Regulatory Challenges and Shortcomings in Asian shore

Asian magnitude of consuming class stirred robust digital revolution and online platforms play a prominent role in the creation of digital value that forms the current and future economic growth (in reshaping its business eco-system. Asian digitalisation process has triggered potential benefits and dangers, mixed with unpredictable and blurred algorithmic intervention, with difficulty to find the fault (Ramaiah, Sirait & Smith, 2019). Although Asian markets have adopted and benefitted the digitalization features in almost all sectors of the economy, with disruptive policies, but at what cost and whose term of development? The 'big data' syndrome has peaked digital corporations to strategically merge for monopoly. M&A is most prominent in rapidly changing technology (especially in high technology) and fierce competition industry like WhatsApp (acquired by Facebook), Grab and Uber. Hence, regulating digital merger become pivotal focus of concern of national and regional CL authorities (Ramaiah,2020). Therefore, digital dividends (development benefits) only reaped if better regulations, human capital, and good governance placed to manage it (Maria, 2017) against the risks of distortionary and adverse effects of digital technologies. Thus, digital technologies to benefit everyone everywhere, need "analog complements" to ensure competition among businesses, by adapting workers' skills to the demands of the new economy, and by ensuring that institutions are accountable (World Bank,2016). But comparatively developing Asia's regulatory environment for competition is weak compared to the scale economies enabled by the internet and digital technologies giant to concentrate their economic power globally. Such network effects can promote dominant entity to grow stronger toward a single, winner-take-all standard in the market (Liu, et al.,2012) and "to act as private regulators setting the rules of the game on the markets they control." and "act as private gatekeepers to critical online activities for an exceptionally large population of private and business users." (Geradin, 2020). The impact is more serious on Asian market because of our borrowed technology dependence.

3.2 Challenges in Using Competition Law on Digital Market in Asia: Case study from Malaysia, Indonesia and India

Grab surfaced as the nationwide and region-wide monopoly in E-hailing industry, from consumer transportation to food delivery and more. (Ramaiah, 2020) when Grab Holdings Inc., and Uber Technologies Inc. merged in 2018. Their horizontal anti-competitive merger unfolded like a tell-tail without any specific regulatory clearance or public disclosure of information and took off unguarded in most SEA nations (Ramaiah, 2020). Singapore, Competition and Consumer Commission of Singapore (CCCS) first to declare Grab-Uber marriage as anti-competitive merger (Section 62 Competition Act 2004 (Chapter 50B)(SCA, (CCCS, 2018) that creates an exclusivity arrangement that would potentially burden new entrant to spend a lot of money to build up driver and rider networks similar in scale and size to the incumbents. The CCCS's decision caused and triggered other SEA government's regulators, mainly Malaysia, Indonesia, Vietnam, Philippines, and Thailand to monitor Uber-Grab merger repercussions on their shore. (Reuters, 2018, 27 July; Ramaiah,2019).

Malaysia Competition Commission(MyCC) although was instructed (The Edge, 2018; The Malaysian Insight, 2018) but was unable to nab the merger transaction for possible legal violations (Khuen, 2018) because

the Malaysian Competition Act 2010 (MCA2020) excluded M &A control and MyCC unable to conduct a market study to determine its pre-merger or post-merger impact on the competition and consumers welfare (Ramaiah, 2020). However, 2019, MyCC's found Grab abused its dominant position by imposing restrictive clauses on its drivers (disallowing competitors advertising) that undermine new entries and other small enterprises to enter the E-haling industry in Malaysia. MyCC proposed a fine of RM86.8 million against Grab for (Jay&Antara, 2019) for abusive behaviour (Section 4 MCA2010) but the decision was set aside and granted leave for judicial review. The case displayed the lack of the regulator's capacity building and legal tools to nab digital merger. And reflected difficulty to prove modus operandi in digitalized based firms because lacking the expertise in EA, SAD to gather the information for the legal proceedings against digital giants such as Grab. Thus, MyCC, national competition watchdog's *parens patriae* was undermined with no merger control(Ramaiah,2020) and lack of expertise in digital case.

In Indonesian's Antimonopoly and Unfair Business Competition (Law No.5/1999) found lex imperfecta, despite having merger control regulation, had to forgo Uber-Grab digital merger (Ramaiah, Sirait & Smith, 2019). The decision queried, whether Indonesian Competition Commission(ICC) lack of regulation or lack the legal tool to assess the impact towards digital competition in Indonesia? ICC attempted to overcome the regulatory gap by introducing 'pre-notification procedure' (Guidelines to measure M&A) for better process in future (KPPU, 2019) but again Gojek-Tokopedia merger took off reflecting the inherent weakness and lack of capacity building in identifying and regulating digital M&A in Indonesian. Meanwhile on later development PT Solusi Transportasi Indonesia (or 'GRAB' Indonesia) was nabbed (KPPU,2020July 02) for the violation Law No.5/1999 for abusive charges 'special rental' amounting to discrimination by monopoly for Grab App software application (KPPU, 2020, September) (Nugraha, 2020). But the decision got reversed by the Supreme Court of Indonesia for reasons of insufficient evidence and penalties did not reflect an effort to economic recovery in the middle of the pandemic. Hence, the decisions were critiqued as a victory fort for digital firms' atrocity and undermined the ICC. And it impairs the competition spirit in Indonesia (Fauzie,2021) because it undermines new and small players (which may go unnoticed) development, and survival compared to the dominant players better equipped digital technology and investment, besides imposing switching costs for people in sharing economy to move from one platform to another (Safiri, 2021).

Meanwhile, technology companies in India, although revered for their innovation and efficiency, also not susceptible to anticompetitive motivated collusion, acquisition, and abuse of market power. Indian digital giants such as Uber, Ola and PayTM in India, market power marginalise the market by enticing users by subsidising their goods with their huge financial capital, by resorting into practices like deep discounting, cash-back offers and other schemes designed to attract new users and establish the network effect (Graham & Smith, 2014) initially even if sustain heavy losses for years. (Chakravartti & Mundle, 2017). Digital markets employ algorithms to limit competition through agreements, concerted practices, and other subtle means. Such as disguising the collusive agreement as an introductory offer by a new player, instead of appearing as systematic competitive strategies by using capital as their competition weapon, where eventually tipping in favour of the player, which might not have the most innovative product or service, but the one that can manage to obtain huge capital and entice as many users as possible through its introductory offers. The Indian, Competition Act, 2002 (ICA2002) and the Competition Commission of India (CCI) similarly faced with difficulty to nab for infringement, in the absence clear evidence to show agreement between platforms to coordinate or be any part of such agreement. Such as between drivers themselves, to delegate the pricing power to the platforms or cab aggregators (Peter & Singh, 2019). Meanwhile "abuse of dominance" cases on global technology giants like Google, Amazon, and their Indian counterparts like Flipkart (CCC,2014: CCi,2019), decisions heavily influenced by European Commission (EC) or of United States FTC/ DOJ decisions per se (Gouri, 2021). However, CCI in WhatsApp's privacy policy and mandatory terms of service that gave users the option to optout decided that "…in a data driven ecosystem, the CL needs to examine whether the excessive data collection and the extent to which such collected data is subsequently put to use or otherwise shared have anti-competitive implications which require anti-trust scrutiny…" (CCI, 2021). Hence, CCI have similar issues in

addressing their market size and algorithm and big data monopoly used or abused with respect to finding infringement on digital based structure.

4. Recommendation and Conclusion to the Legal Conundrum in Digital Economy

A nuanced regulatory approach much required to balance innovation and to focus on platform markets usage of algorithm to set prices. In the war between Man versus system, sovereignty and efficiency are beginning to show strong potential for conflict on the following issues: the nature and protection of data, the way it is communicated and stored, the models and algorithms according to which it is processed and used the authority it is given to act and the people whose labour it replaces. The Malaysian, Indonesian and Indian case study reflects the enforcement conundrum in Asia's digital economy market because far ill-equipped, need more political will as well as expertise and experts to deal with the digital giants. Asian regulators digital management and choice of direction on digital market mindsets also more protective of the competitor to avoid the potential disincentive effect of cooling on the degree of innovations that may restrict the digital economy. The local Asian regulator needs to embed expertise and expert tools as part of their legal fact-finding to identify the perpetrator in the digital plethora. Therefore, competition agencies must enhance their existing procedure by appointing specialists in digital markets units and officers to catch-up with the fast-paced digital markets technological development. And traditional tools for competition analysis must be refinement to address better the specificities of online markets, such as the multisided nature of platforms, network effects, zero-price markets, so-called 'big data' and the increased use of algorithms to find the right balance between the man and not anymore, the machine but the digital embedded brain, the algorithm, AI, or big-data owner to unveil the perpetrator behind the screen.

Acknowledgements

Authors are grateful and thankful to University Technology MARA (UiTM), Pulau Pinang for all the financial assistance given for this conference research paper.

References

Case Number 13/KPPU-I/2019 on July 2, 2020
CCI (2014) Mohit Manglani Vs. Flipkart & Ors Competition Commission of India, Case No.80 of 2014 Retrived from https://www.cci.gov.in/sites/default/files/802014.pdf
CCI (2018) Competition Commission of India, Case No.20 of 2014. Retrived from https://www.cci.gov.in/sites/default/files/20-of-2018.pdf
CCI (2019) In All India Online Vendors Association Vs. Flipkart & Ors Case No. 40 of 2019, Competition Commission of India. Retrieved from https://www.cci.gov.in/sites/default/files/40-of-2019.pdf
CCI (2020) Competition Commission of India, Case No.15 of 2020, para no.101. Retrived from https://www.cci.gov.in/sites/default/files/15-of-2020.pdf
CCI (2021) Competition Commission of India, Suo Mote Case No.1 of 2021, para 30 - Retrived from http://www.cci.gov.in/sites/default/files/SM01of2021_0.pdf
CCI 2016) Case No. 99 of 2016Whatsapp, Competition Commission of India, Case, Retrieved at Retrived fromhttps://www.cci.gov.in/sites/default/files/26%282%29%20Order%20in%20Case%20No.%2099%20of%202016.pdf?download=1
CCI Releases (2020, January 8) 'Market Study on E-commerce in India: Key Findings and Observations'. Retrived from https://www.cci.gov.in/sites/default/files/whats_newdocument/Market-study-on-eCommerce-in-India.pdf
Chakravartti,P. & Mundle,S. (2017) An Automatic Leading Indicator Based Growth Forecast For 2016-17 and The Outlook Beyond. National Institute of Public Finance and Policy New Delhi (NIPFP Working Paper

No. 193 30-Mar-2017. Retrieved from
http://www.nipfp.org.in/media/medialibrary/2017/04/WP_2017_194.pdf
Competition Act 2010, Malaysia
Competition Act, 2002, India as amended by the Competition (Amendment) Act, 2007 ICA (2002) Retrieved from https://www.cci.gov.in/competition-act
EA,Evidence Algorithm. Retrieved from http://www.nevidal.org/
EU (2017, May10) Final report on the E-commerce Sector Inquiry, European Commission p. 4.
Fauzie, Y.Y (2021, May, 18) Membaca Tujuan dan Dampak Merger Gojek-Tokopedia.CNN Indonesia. Retrieved from https://profesiakuntanpublik.com/membaca-tujuan-dan-dampak-merger-gojek-tokopedia/ |
Foda, K.& Patel, N. (2018) Brookings. Competition challenges in the digital economy. Retrieved from https://www.brookings.edu/blog/up-front/2018/06/28/competitionchallenges-in-the-digital-economy
Fung, Brian (2020, October 10) Near-perfect market intelligence': Why a House report says Big Tech monopolies are uniquely powerful. CNN Business.
Geradin, D.(2020, October 5)What is a digital gatekeeper? The Platform Law Blog. Retrieved from https://theplatformlaw.blog/2020/10/05/what-is-a-digital-gatekeeper/
Gouri, G & Salinge,M (2017) Protecting Competition v/s Protecting Competitor: Assessing the Antitrust Complaints against Google, , The Criterion Journal of Innovation, Vol 2, 2017, p 531-558.
Gouri, G (2021) Platform Markets – The Antitrust Challenge in India, CPI.Retrieved from https://www.competitionpolicyinternational-com.opj.remotlog.com/platform-markets-the-antitrust-challenge-in-india/
Graham,C. & Smith,F. (2014) (eds.) "Competition, Regulation and the New Economy, p. 17 to 53. Hart Publishing
Jay, B. T. & Antara, B. (2019, October 3) Malaysia proposes RM86 million fine on Grab for abusive practices. New Straits Times. Retrieved from
https://www.nst.com.my/news/nation/2019/10/526697/malaysia-proposes-rm86-million-fine-grab-abusive-practices
Khuen, L. W. (2018, April 10). The Sun. MYCC assessing impact of Uber–Grab merger. Retrieved from http://www.thesundaily.my/news/2018/04/10/mycc-assessing-impact-uber-grab-merger
KPPU (2020, July 2) Case Number 13/KPPU-I/2019
KPPU (2020, September) ICC will file a cassation over Grab Case .Retrieved from https://eng.kppu.go.id/icc-will-file-a-cassation-over-grab-case/.
KPPU(2019) Peraturan Komisi Nomor 3 Tahun 2019) Retrieved from https://kppu.go.id/wp-content/uploads/2020/10/Pedoman-Penilaian-Terhadap-Penggabungan-Peleburan-Atau-Pengambialihan-FINAL.pdf.
Law No.5/1999, Indonesia Antimonopoly and Unfair Business Competition Act
Liu, C.s Z., Kemerer, C. F. Slaughter, S. and Smith, M. D. (2012), Standards Competition in the Presence of Digital Conversion Technology: An Empirical Analysis of the Flash Memory Card Market (July 1, 2012). MIS Quarterly,
Maria. R.S., Urata,S and Intal, P.S.Jr. (2017)The Integrative Chapter: The ASEAN Economic Community Into 2025 and beyond. Volume 5 | The ASEAN Economic Community Into 2025.ERIA.
Marz,S., Cerf,Moran & Rolnik, G. (2018, April 10) Solutions to the Threats of Digital Monopolies. Promarket, Stigler Centre, University of Chicago. Rerieved from https://promarket.org/2018/04/10/solutions-threats-digital-monopolies/
Microsoft (2019) Hub and spoke network topology. Retrieved from https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/ready/azure-best-practices/hub-spoke-network-topology.
News Release (2021, February 4) Senator Klobuchar Introduces Sweeping Bill to Promote Competition and Improve Antitrust Enforcement. Retrieved from
https://www.klobuchar.senate.gov/public/index.cfm/2021/2/senator-klobuchar-introduces-sweeping-bill-to-promote-competition-and-improve-antitrust-enforcement
Nugraha, R. M. (2020 July 3) Court Fines Grab Indonesia over Illegal Business Practices. Tempo. Co. Retrieved from https://en.tempo.co/read/1461666/indonesian-startups-targeting-public-investors-with-ipo
Peter, A & Singh, Neha (2019) The Evolution of Competition Law in Digital Markets in India. CPI. Antitrust Chronicle, Vol 1(1).
Ramaiah, A. K., Sirait, N. N., & Smith, N. N. (2019). Competition in Digital Economy: The State of Merger Control on Consumer Transportation in ASEAN. International Journal of Modern Trends in Business Research (IJMTBR), 2(7), 66-82.

Ramaiah, A.K. (2020). Merger Phenomena In Digital Economy: Uber-Grab Competition Tell-Tale In Malaysia. EpSBS - Volume 88 - AAMC 2019 doi:10.15405/epsbs.2020.10.56
Reuters (2018, 27 July) Grab defends position in Uber deal to Singapore's anti-monopoly watchdog. Retrieved from https://www.reuters.com/article/us-uber-grab-singapore/grab defends-position-in-uber-deal-to-singapores-anti-monopoly-watchdog idUSKBN1KH067.
Ritter, E & Solyom, I. (2018) Hungary: Competition And Consumer Law Enforcement In Digital Markets In Hungary. Mondaq.
Safiri, K. (2021, May 24)Merger dengan Tokopedia, Ini Manfaat yang Diterima Gojek Kompas.com. Rerieved from https://money.kompas.com/read/2021/05/24/155130926/merger-dengan-tokopedia-ini-manfaat-yang-diterima-gojek.
The Edge (2018, 13 July) MyCC to probe Grab-Uber merger. Retrieved at http://www.theedgemarkets.com/article/mycc-probe-grabuber-merger
UNCTAD (2010) Model Law on Competition, United Nation Conference on Trade and Development. Retrieved at https://unctad.org/system/files/official-document/tdrbpconf7d8_en.pdf
Whish, R., & Bailey, D. (2015). Competition Law (Eighth Edition). Oxford University Press.
World Bank, 2016. Retrieved from https://www.worldbank.org/content/dam/Worldbank/Publications/WDR/WDR%202016/WDR2016_overview presentation.pdf

Development of Down Syndrome Child Assessment Application Prototype

Syahrul Mauluddin[a*], Marliana Budhiningtias Winanti[b], Dadang Munandar[c], Imelda Pangaribuan[d], Feisal Abdurrahman[e], Muhamad Chairil Akmal[f]

[a,b,d,e,f]Universitas Komputer Indonesia, Jl. Dipatiukur No. 102-116, Bandung and 40132, Indonesia
[c]Universitas Wanita Internasional, Jl. Pasir Kaliki No. 179 A Bandung 40173, Indonesia
* Email: syahrul.mauluddin@email.unikom.ac.id

Abstract

This study aims to develop a prototype of assessment application for children with Down syndrome. This application is to help parents with Down syndrome children in detecting the growth and development of Down syndrome children and making development programs. The development of this application is motivated by the fact that children with Down syndrome have various weaknesses in physical, health and IQ. These weaknesses cause a high risk of developmental delays in children with Down syndrome. Therefore, every parent with Down syndrome child needs to know each stage of their child's development then carry out a child development program to prevent delays in their child's development. In developing the assessment application for children with Down syndrome, a prototype model system development method is used with an object-oriented approach. Activities in this research are database design, interface design, and build assessment of children with Down syndrome applications. Presence od Down syndrome child assessment application, hopefully can help parents of Down syndrome children in detecting the development of their child's development so that it is known whether or not there is a delay in their child's development. In addition, parents can also make a child development program followed by making a daily record of the process and results of implementing their child's development program, so that their child's growth and development can be monitored and become evaluation material for further development programs.

Keywords: Application, Down Syndrome, Assessment, POTADS;

1. Introduction

Children are the greatest gift from God to human being. In creating the human child, God has his own secret. Some children are born normal, and some are born special. One of them is a child with Down syndrome. Down syndrome is a genetic disorder that causes people to experience delays in growth, as well as physical disabilities and weaknesses. The number of children with Down syndrome is born in various parts of the world, according to the World Health Organization (WHO), the birth rate with Down syndrome is 1: 1,000 births worldwide. Each year, about 3,000 to 5,000 children are born with Down syndrome (InfoDATIN Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia, 2019).

Congenital or birth defects such as Down syndrome cannot be treated, so the best effort for parents is to provide proper parenting so that children with Down syndrome can live independently. Because the main task faced by someone with a disability is to achieve independence (Hasanah, Wibowo, & Humaedi, 2016). The lack of knowledge of the parents about the care and fulfillment of child with down syndrome (CDS) rights causes CDS to adulthood not having independence, from caring for themselves to carrying out their social functions.

To help parents understand how to treat children with Down syndrome, some CDS parents join the Parents Association of Children with Down Syndrome (POTADS) community. The purpose of POTADS was formed to empower parents who have children with Down syndrome to be always eager to help their special child grow and develop (POTADS, 2020).

To help the growth and development of CDS optimally, the abilities needed by CDS parents are to assess and compile an CDS development program, so that parents can evaluate CDS weaknesses and improve CDS

abilities programmatically / directed from an early age. However, the fact is that the expertise of this assessment is generally only owned by people / institutions in the field of therapy and inclusive education. Assessment is a process of determining the results that have been achieved by several planned activities to support the achievement of objectives (Arikunto & Jabar, 2008).

2. Literature Review

Down syndrome is a disorder that causes people to have various weaknesses such as physical (weak muscles), health (disabilities and susceptibility to disease) and low IQ. Slow development is a major feature of Down syndrome children (Mulia & Kristi, 2012).

In addition, people with Down syndrome have distinctive facial features, including folds at the corners, slanted eyes that tend to point upwards, a flat nose, a face that looks like a snout, and a small mouth with a flat palate so that their tongue sticks out a little (Rachmawati & Masykur, 2016).

This research was conducted inseparable from the results of previous studies that had been conducted. Based on the results of research conducted by Christine Leontia and Nina Sevani entitled "Web for Early Detection of Down Syndrome Retardation in Children". The application design focuses on determining the rate of retardation in children, which consists of mild, moderate and severe retardation. Input is made in the form of user answers regarding attitudes and behavior in children to determine the level of retardation in children, and the resulting output is the level of retardation for children with Down syndrome, as well as suggestions for further treatment for children with Down syndrome (Leonita & Sevani, 2015).

While in this research, the application design focuses on assessment, evaluation and development programs for children with down syndrome (CDS). Through the AADS application, CDS parents can make CDS development plans based on assessment by focus to the child development standards provided in the AADS application. In addition, in each program, CDS parents are required to input daily notes as a realization of the CDS development plan. For parents, the output generated from this application is in the form of a history of planning and achievements of the CDS development program, while for administrators, the output produced is in the form of information from all the planning and achievements of CDS development programs carried out by CDS parents which can then be used as a reference for growth and development data of CDS.

3. Method

In developing assessment of children with Down syndrome application using an object-oriented approach and a prototype model system development method. The prototype model used is as shown in Figure 1.

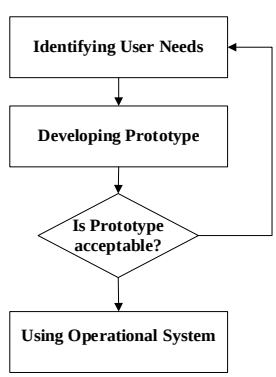


Fig. 1. Prototype Model (McLeod & Schell, 2004)

The explanation of each stage is as follows:

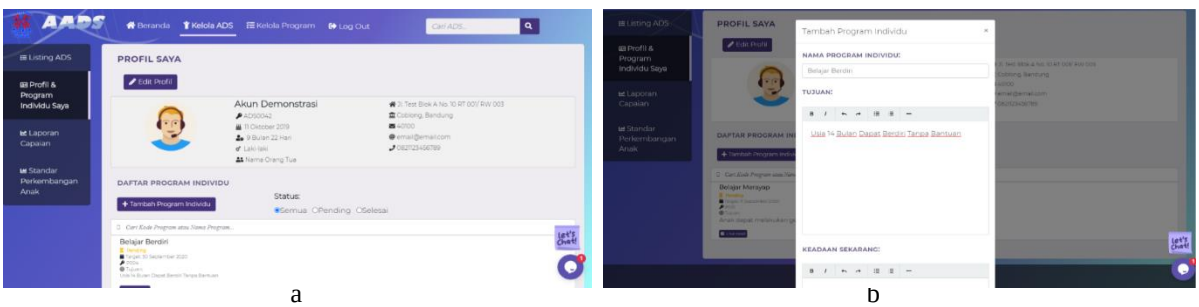
1. Identifying user needs. The first stage is to identify the problem and find a solution.
2. Developing prototype. The second stage is to design and build a prototype application for the Down Syndrome Child Assessment including database design, interface design and coding.
3. Determine whether the prototype is acceptable. The third stage is testing by user. If there is still something that needs to be fixed then go back to the first stage.

In developing the application prototype, there is a testing process carried out by the black box method. Black box testing is software quality testing that focuses on software functionality (Hidayat & Muttaqin, 2018) (Mustaqbal, Firdaus, & Rahmadi, 2015).

4. Result

Based on the result of system requirements analysis, this down syndrome child assessment application has nine functions, namely: registration, profile management, individual programs management, add daily notes, print outcome report, browse child development standards, article management, child development standards management, and chat (Mauluddin, Winanti, & Munandar, in Press).

At development stage, it is divided into some activities that are designing databases, designing interfaces and coding. Based on the results of database analysis and design, the AADS application database consists of 6 tables, namely: user, program, child_development, daily_notes, development_standards and article. The results of interface design and coding are as shown in Fig 2a and Fig 2b.



a b

Fig. 2. (a) Profil Management; (b) Individual Programs Management

Fig 2a is a feature to management of profile and Fig 2b is a feature for creating Down syndrome child development programs by inputting data on individual program names, program objectives, current conditions, source materials / props, target dates and development categories.

At the testing stage, the testing method used is black box. Based on the test results of 9 features in the AADS application, all features can run well and according to the expected functions.

5. Conclusion

With the Down syndrome child assessment application, it is expected that parents of Down syndrome children can be helped in detecting the development of their child's development so that it is known whether or not there is a delay in their child's development. In addition, parents can also make their child development program followed by keeping a diary of the process and results of implementing their child's development program, so that their child's growth and development can be monitored and become evaluation material for further development programs.

Acknowledgements

Authors would like to thank the ministry of education, culture, research and technology for funding this research in 2021 Fiscal Year.

References

Arikunto, S., & Jabar, C. S. (2008). *Evaluasi Program Pendidikan: Pedoman Teoritis Praktisbagi Mahasiswa dan Praktisi Pendidikan.* Jakarta: Depdiknas.

Hasanah, N. U., Wibowo, H., & Humaedi, S. (2016). Pola Pengasuhan Orang Tua dalam Upaya Pembentukan Kemandirian Anak Down Syndrome. *Share Social Work Journal, 5*(1), 65-70.

Hidayat, T., & Muttaqin, M. (2018). Pengujian Sistem Informasi Pendaftaran dan Pembayaran Wisuda Online menggunakan Black Box Testing dengan Metode Equivalence Partitioning dan Boundary Value Analysis. *Jurnal Teknik Informatika UNIS (JUTIS), 6*(1), 25-29.

InfoDATIN Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia. (2019). *Antara Fakta dan Harapan Sindrom Down.* Kemenkes RI.

Leonita, C., & Sevani, N. (2015). Web Untuk Deteksi Dini Tingkat Retardasi Down Syndrome Pada Anak. *Jurnal Teknik Informatika dan Sistem Informasi, 1*(1).

Mauluddin, S., Winanti, M. B., & Munandar, D. (in Press). Analysis of System Requirements of Children with Down Syndrome Assessment Application. *Journal of Physics: Conference Series*.

McLeod, R., & Schell, G. (2004). *Sistem Informasi Manajemen.* Jakarta: Indeks.

Mulia, A., & Kristi, E. (2012). Fasilitas Terapi Anak Down syndrome di Surabaya. *JURNAL eDIMENSI ARSITEKTUR,*(1), 1-6.

Mustaqbal, M. S., Firdaus, R. F., & Rahmadi, H. (2015). Pengujian Aplikasi Menggunakan Black Box Testing Boundary Value Analysis (Studi Kasus : Aplikasi Prediksi Kelulusan SNMPTN). *JITTER, 1*(3), 31-36. Retrieved from https://adalah.co.id/white-box-testing/

POTADS. (2020). *Tentang Kami*. Retrieved from POTADS: https://potads.or.id/tentang-kami/

Rachmawati, S. N., & Masykur, A. M. (2016). Pengalaman Ibu yang Memiliki Anak Down Syndrome. *Jurnal Empati, 5*(4), 822-830.

Issues in Surgical Scheduling Problem: Uncertainty, Capacity Planning, Request and Demand

Norizal Abdullah[a*], Masri Ayob[a], Meng Chun Lam[b], Nasser R. Sabar[c]

[a]Data Mining and Optimization Lab, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
[b]Mixed Reality and Pervasive Computing Lab, Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia Bangi Selangor, Malaysia
[c]Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
* Email: norizal.abdullah23@gmail.com

Abstract

The surgical scheduling problem in healthcare is one of the most crucial problems that get attention from many researchers. However, less concern is given to understand the issue in the surgical scheduling problem. The purpose of this review is to reveal the crucial issues in a surgical scheduling problem. We organize the issues into three categories: (1) Uncertainty; (2) Capacity Planning; and (3) Request and Demand. In the uncertainty issue, we discussed the issue that may cause uncertainty that happens in surgical procedure, and for capacity planning issue, we concentrate on resources planning such as human and material resources. The request and demand issue is related to the surgical case criteria and surgical team preferences. At the end of the review, we determine the important issue that arises inside the surgical scheduling problem requires further attention.

Keywords: Healthcare; Surgical Scheduling Problem; Hospital

1. Introduction

The surgical scheduling problem is described as the selection of procedures to be performed, the allocation of resource time to those procedures, and the sequencing of those procedures within the time allotted (May, Spangler, Strum, Vargas, & Management, 2011). (Belkhamsa, Jarboui, & Masmoudi, 2018) state that pre-operative, intra-operative, and post-operative phases are stages in a surgical procedure. The surgery would require both material and human resources to complete the procedures. Pre-operative Holding Units (PHU) beds and nurses were required during pre-operative. Next, the intra-operative stage needs to consider the Operating Room (OR), surgeons, anaesthetists, and nurses. Finally, services such as Incentive Care Unit (ICU) beds and nurses are required for the post-operative phase.

There are two types of surgery: elective and emergency surgery (Silva & de Souza, 2020). Under elective surgery, three types of surgery schedule strategy can be used either block scheduling, open scheduling and modified block scheduling (Patterson, 1996). Although there is guidance for the surgery procedure and various type of surgery schedule strategy to be used, there are still issues that may cause the surgical scheduling disruption. (May, Spangler, Strum, Vargas, & Management, 2011), state that disruptions that happen in surgical scheduling are mainly due to the uncertainty issue from the procedure duration and capacity planning issue. The unbalanced scheduling of the OR department also often causes demand fluctuation in other departments such as surgical wards and intensive care units (J. M. van Oostrum, Van Houdenhoven, Hurink, Hans, Wullink, & Kazemier, 2008). (J. van Oostrum, van Houdenhoven, Wagelmans, & Kazemier, 2009) propose Master Surgical Scheduling (MSS) to deal with this issue. The MSS implementation improved coordination among hospital staff members and solve the issue of long-term forecasting and capacity planning. Another advantage is that it boosts OR performance and eliminates the problem of confusion. (Silva & de Souza, 2020) proposed an optimization method for the ambiguity problem in surgical scheduling by incorporating approximate dynamic programming. Their method has the potential to make a big difference in practice. Resource constraints in Surgical Department also may lead to the issue in surgical scheduling. As example, (Wang & Xu, 2017)

proposed an evolutionary algorithm to deal with the issue of multiple resources constraints in their study. Fig. 1. show the surgical procedure phase (Belkhamsa, Jarboui, & Masmoudi, 2018). We illustrate the part the surgical scheduling problem may happen that relate to an issue that already states previously.

2. Issue Representation

From the analysis of previous studies, we categories the issue in surgical scheduling problem into three types which are: (1) Uncertainty; (2) Capacity Planning; and (3) Request and Demand. In this section, we discuss the

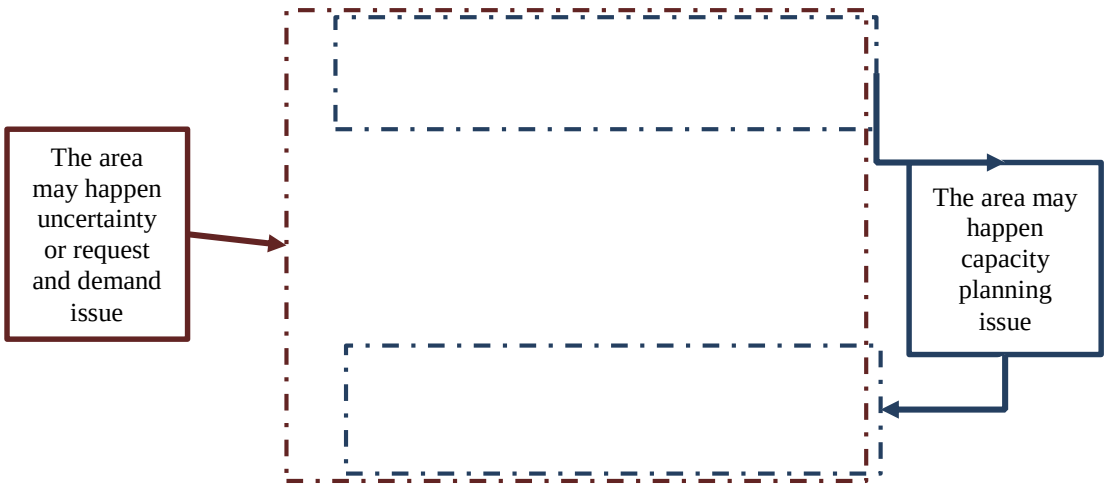


Fig. 1. Surgical procedure phase

uncertainty issue that commonly arises in surgical scheduling. We then address the capacity issue that relates to resources planning, and subsequently, the problems of request and demand related to the surgical case criteria and surgical team preference.

2.1 Uncertainty

Uncertainty is the most crucial issue that arises in surgical scheduling. Surgery duration is one of the reasons that cause the uncertainty issue (J. M. van Oostrum, Van Houdenhoven, Hurink, Hans, Wullink, & Kazemier, 2008). The surgery's duration is unknown due to the procedure's unforeseen start time and leads to an increase in overtime (Zhang, Wang, Tang, & Lim, 2020). Overtime is a frequent occurrence in surgery units, resulting in increasing surgery team or patient stress or dissatisfaction and financial loss for hospitals (Zhang, Wang, Tang, & Lim, 2020). (Silva & de Souza, 2020) discussed the uncertainty on when emergency and urgent surgery will arrive. When emergency and urgent surgery occurs without warning or intervention, the confusion about its arrival will be exposed.

Furthermore, the uncertainty issue also may happen due to the recovery time for the patient. For example, (Zhu, Zhang, Jiao, & Li, 2015) state uncertainty may arise due to the patient's recovery time by the anaesthesia after the surgery. (Wiyartanti, Park, Chung, Kim, Sohn, & Kwon, 2015) stated that when many surgical cases need immediate changes in data, such as when surgery delays or cancellation occurs on the same day, confusion can arise from the patient or the surgical procedure itself.

2.2 Capacity Planning

(Chow, Puterman, Salehirad, Huang, Atkins, & Management, 2011) study the way to improve the surgical scheduling caused by the high surgical bed utilisation. Because of the capacity problem, this situation is

challenging to be handled. The issue occurs due to the demand peaks, insufficient space, a lack of capacity in speciality-specific surgical wards, and available beds in surgical wards. Another capacity issue that should be considered is recovery bed capacity in the surgical department. (Astaraky & Patrick, 2015) state that the recovery bed capacity should be assured such that the available beds of recovery is adequate.

The limitation of human and material resources, such as nurses, auxiliary personnel, medical equipment, or places in intensive care units, is also a factor in the capacity problem (Silva, de Souza, Saldanha, & Burke, 2015). Another capacity issue is related to the limitation of the operating room (Liu, Wang, & Wang, 2019).

2.3 Request and Demand

Request and demand are common in open scheduling, which allows the surgeon to select their favourite surgical day (Dexter, Traub, & Macario, 2003). When the patient requests the surgery date, another request and demand will occur (Silva & de Souza, 2020). The request often occurs when a surgical request is sent from the hospital's waiting list (Banditori, Cappanera, & Visintin, 2013). Frequent request and demand change certainly cause the surgical scheduling problem.

(Yang, Shen, Gao, Liu, & Zhong, 2015) studied surgeon demand in surgical scheduling based on allocating the right surgeon in the operation room with the matching time resources. Also, the demand issue that may relate to the surgical scheduling problem is the demand for resources. (Wang & Xu, 2017) stated that resource demand is the resources required to complete a surgical procedure. For example, in the intra-operative phase resources such as material and human resources is needed to perform this phase. Besides that, (Guda, Dawande, Janakiraman, Jung, & Management, 2016) stated that the surgeon may sometimes request the patient arrive earlier than the scheduled surgery time.

3. Conclusion

This work reviews the issue of a surgical scheduling problem and categorises the issues into three types which are (1) Uncertainty; (2) Capacity Planning; and (3) Request and Demand. Based on the review, we can conclude that all these issues are interrelated and need to be considered in order to cope with the disruption in the surgical schedule. This ambiguity problem extends beyond the surgical scheduling process and should be considered in designing a good quality surgical schedule that can cope with disruptions. For future work, we would like to investigate other issues in scheduling that relates to the surgery procedure. An example is an issue in scheduling in human resource in surgical procedures such as nurse scheduling, surgeon scheduling, and anaesthetist scheduling.

Acknowledgement

The authors wish to thank the Universiti Kebangsaan Malaysia and the Ministry of Higher Education Malaysia for supporting and funding this work (grant ID: TRGS/1/2019/UKM/01/4/1).

References

Astaraky, D., & Patrick, J. J. E. J. o. O. R. (2015). A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling. *245*(1), 309-319.

Banditori, C., Cappanera, P., & Visintin, F. J. I. J. o. M. M. (2013). A combined optimization–simulation approach to the master surgical scheduling problem. *24*(2), 155-187.

Belkhamsa, M., Jarboui, B., & Masmoudi, M. (2018). Two metaheuristics for solving no-wait operating room surgery scheduling problem under various resource constraints. *Computers & Industrial Engineering, 126*, 494-506. doi:https://doi.org/10.1016/j.cie.2018.10.017

Chow, V. S., Puterman, M. L., Salehirad, N., Huang, W., Atkins, D. J. P., & Management, O. (2011). Reducing

surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *20*(3), 418-430.
Dexter, F., Traub, R., & Macario, A. (2003). How to Release Allocated Operating Room Time to Increase Efficiency: Predicting Which Surgical Service Will Have the Most Underutilized Operating Room Time. *Anesthesia & Analgesia, 96*, 507-512. doi:10.1213/00000539-200302000-00038
Guda, H., Dawande, M., Janakiraman, G., Jung, K. S. J. P., & Management, O. (2016). Optimal policy for a stochastic scheduling problem with applications to surgical scheduling. *25*(7), 1194-1202.
Liu, L., Wang, C., & Wang, J. J. J. o. C. O. (2019). A combinatorial auction mechanism for surgical scheduling considering surgeon's private availability information. *37*(1), 405-417.
May, J. H., Spangler, W. E., Strum, D. P., Vargas, L. G. J. P., & Management, O. (2011). The surgical scheduling problem: Current research and future opportunities. *20*(3), 392-405.
Patterson, P. (1996). What makes a well-oiled scheduling system? *OR Manager, 12*(9), 19-23.
Silva, T. A., de Souza, M. C., Saldanha, R. R., & Burke, E. K. J. E. J. o. O. R. (2015). Surgical scheduling with simultaneous employment of specialised human resources. *245*(3), 719-730.
Silva, T. A., & de Souza, M. C. J. O. (2020). Surgical scheduling under uncertainty by approximate dynamic programming. *95*, 102066.
van Oostrum, J., van Houdenhoven, M., Wagelmans, A., & Kazemier, G. J. I. (2009). Implementing a master surgical scheduling approach in a regional hospital. *39*(6), 549-551.
van Oostrum, J. M., Van Houdenhoven, M., Hurink, J. L., Hans, E. W., Wullink, G., & Kazemier, G. J. O. s. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments. *30*(2), 355-374.
Wang, J., & Xu, R. (2017). *Surgical scheduling with participators' behavior considerations under multiple resource constraints.* Paper presented at the 2017 International Conference on Service Systems and Service Management.
Wiyartanti, L., Park, M.-W., Chung, D., Kim, J. K., Sohn, Y. T., & Kwon, G. H. (2015). *Managing uncertainties in the surgical scheduling.* Paper presented at the MIE.
Yang, Y., Shen, B., Gao, W., Liu, Y., & Zhong, L. J. J. o. C. O. (2015). A surgical scheduling method considering surgeons' preferences. *30*(4), 1016-1026.
Zhang, Y., Wang, Y., Tang, J., & Lim, A. J. O. (2020). Mitigating overtime risk in tactical surgical scheduling. *93*, 102024.
Zhu, Y., Zhang, Y., Jiao, Z., & Li, D. (2015). *Surgical scheduling under patients' uncertain anesthesia recovery time.* Paper presented at the 2015 12th International Conference on Service Systems and Service Management (ICSSSM).

Design of Halal Certification Status Checker Application System Using QR-Code

Hidayat[a]*, Afrizal Imanullah[b]

[a,b] Computer Engineering, Universitas Komputer Indonesia, Dipati Ukur No. 102, Bandung 40132, Indonesia
* Email: hidayat@email.unikom.ac.id

Abstract

This paper proposes to design a system to easily determine the halal certification status of a product circulating in Indonesian society. This is conducted to overcome the anxiety of the Muslim community in Indonesia on halal-labeled products that are sold in shops and supermarkets. The system designed consists of two parts, namely the system using PHP programming for recording the halal certification of products and generating QR codes, and the system for reading the status of the product's halal certification from QR codes using android. The information displayed on android is the name of the product, the product manufacturer and the status of the product's halal certification, and other information. The results show that the application successfully reads the QR code and displays information on the halal certification status of the product. The farthest distance can be done up to a distance of 30 cm in bright light conditions, while in low light conditions it is 26 cm.

Keyword: QR code; halal certificate status; android; PHP programming.

1. Introduction

The circulation of products that are suspected of being illegitimate in the Muslim community has caused unrest in the Muslim community. This problem is also triggered by the difficulty of the public in knowing whether or not the LPPOM MUI halal label printed on the product is true. Generally, every product that has been certified halal by LPPOM MUI will include the halal certificate identity number on the product label. However, it is not easy for people to find the halal certification status of these products because there is no more practical application to find out the halal certification status of these products. Halal certification actually has a validity period, so it is important for the public to easily find out halal certification status to get comfort and peace in consuming and using the product.

This paper proposes to design a system to easily determine the halal certification status of a product circulating in Indonesian society using QR code technology (Tiwari 2017). This is conducted to overcome the anxiety of the Muslim community in Indonesia on halal-labeled products that are sold in shops and supermarkets. QR codes have been widely applied in various applications, one of which is the problem of traceability, such as Kim and Woo (2016), Tarjan et al. (2014) and Zhu and Tang (2015) studies.

2. Design

The system designed consists of two parts, namely the system for recording the halal certification of products and the system for reading the status of the product's halal certification as shown in Fig. 1. First part, the product halal certification recording system is built using PHP programming on a web-based. This system will record the product's halal certification status and generate a QR code with version 1 size as a code for the product's halal certification label. The second part, the system for reading the product's halal certification status is built using Android-based Java programming. The QR code accommodates the product halal certificate number from LPPOM which is generated from the product halal certification recording system. Furthermore, the Android application will read the QR code so that it can be read directly by the system without entering the number by the number on the certificate. After that, the number will be matched with the information in the product halal

certification recording system. Further, the information found will be displayed on the android smartphone screen. The information displayed is the name of the product, the product manufacturer and the status of the product's halal certification, and other information.

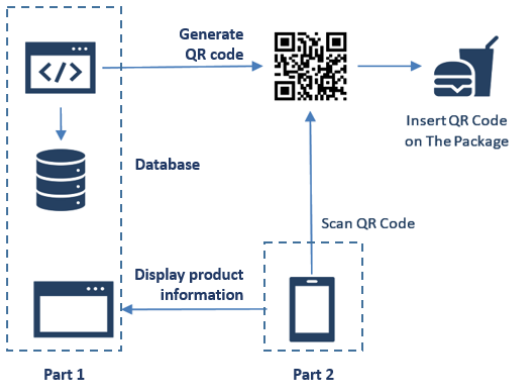


Fig. 1. The block diagram of halal product reading application

3. Result

The product halal certification recording system has two main menus, namely Data users and Data product menus as shown in Fig. 2. It is used to input data by an admin. Admin will enter a list of users who can operate this system. And also admin can input data products that have certificate ID. Fig. 3 and Fig. 4 show displays of data user menu and data product menu respectively.

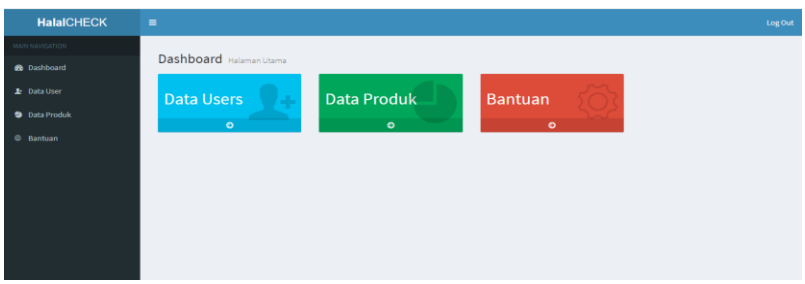


Fig. 2. Display of dashboard menu

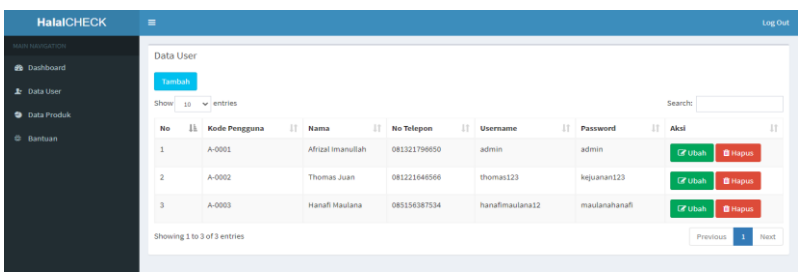


Fig. 3. Display of data user menu

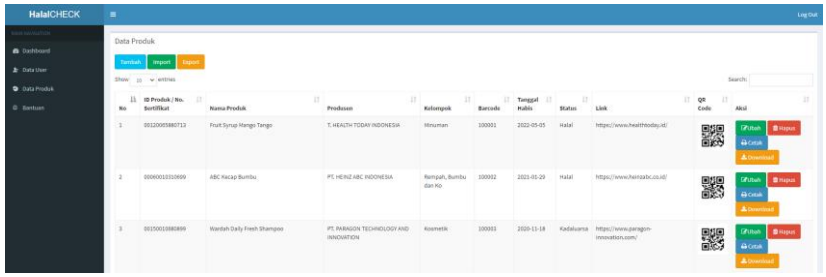


Fig. 4. Display of product data menu

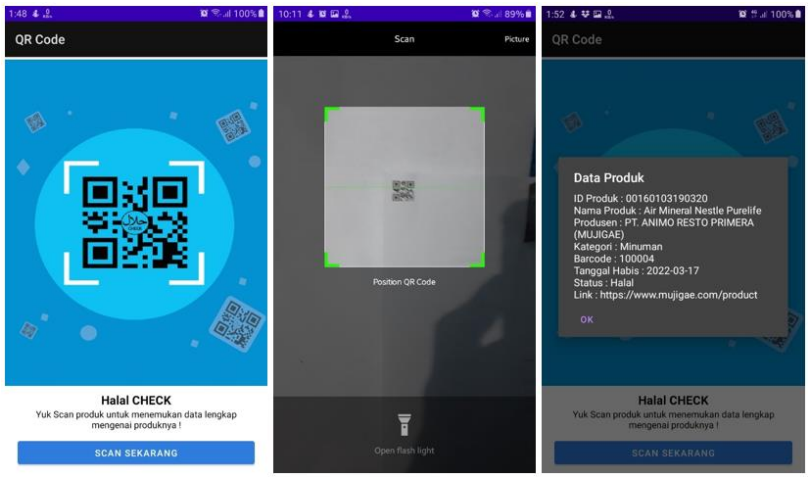
This study uses 100 sample data with product halal certification status sourced from LPPOM MUI attachment of halal certification data. This data is used to test the designed system whether can function properly or not. Testing of halal status certification checker applications is conducted on several QR codes that are printed with a size of 70 pixels. Android App Scanning is conducted at different distances. The results show that the application successfully reads the QR code and displays information on the halal certification status of the product as shown in table 1 and Fig. 5. Fig. 6 shows the reading QR code by the android system. The furthest reading distance using the android system in bright light conditions is 30 cm, while in low light conditions it is 26 cm.

Table 1. Testing of five QR code reading example

No	Product ID	QR-Code	Result
1	00120065880713		[] success
2	00060010310699		[] success
3	00150010680899		[] success
4	00160103190320		[] success
5	152000044220120		[] success
6	00250041070706		[] success
7	00250084330817		[] success
8	01011037630309		[] success
9	17010055420720		[] success
10	00240082120417		[] success

id_produk	nama_produk	produksi	barcode	tanggal_habis	status	kategori	link
00120065880713	Fruit Syrup Mango Tango	T. HEALTH TODAY INDONESIA	100001	2022-05-05	Halal	Minuman	www.healthtoday.id
00060010310699	ABC Kecap Bumbu	PT. HEINZ ABC INDONESIA	100002	2021-01-29	Halal	Rempah, Bumbu dan Ko	https://www.heinzabc.co.id/
00150010680899	Wardah Daily Fresh Shampoo	PT. PARAGON TECHNOLOGY AND INNOVATION	100003	2020-11-18	Kadaluarsa	Kosmetik	https://www.paragon-innovation.com/
00160103190320	Air Mineral Nestle Purelife	PT. ANIMO RESTO PRIMERA (MUJIGAE)	100004	2022-03-17	Halal	Minuman	https://www.mujigae.com/product
152000044220120	AIR MINERAL AQUA	PPBISMG DEPASTRY HOMEMADE	100006	2022-01-21	Halal	Minuman	
00250041070706	Delisla Strawberry Jam	PT Bintang Jaya Baharriski	100007	2020-09-04	Kadaluarsa	Selai dan Jelly	www.bintangjb.com
00250084330817	Strawberry Jam	Ltd Green Juice (Tianjin) Co.	100008	2021-09-17	Halal	Selai dan Jelly	
01011037630309	Rolade Ayam-Choice L	CV. Fiva Food & Meat Supply	100008	2021-05-28	Halal	Daging dan Produk Da	https://www.fivafood.com/
17010055420720	Dendeng Balado Uni Etty	Dendeng Balado Uni Etty	100009	2020-07-22	Kadaluarsa	Daging dan Produk Da	
00240082120417	Ektrak Nabati	CV. Ocean Fresh	100010	2021-03-31	Halal	Ekstrak	http://www.oceanfresh.id/

Fig. 5. List of data product information



(a) (b) (c)

Fig. 6. Halal CHECK android application, (a) Main display, (b) QR code scanning, and (c) Information halal product

4. Result

The system that is built can function properly. Part one of the system can store product data that has a halal certificate and generate a unique QR code per product ID. In addition, the second part of the system can read the QR code and display product information including the validity period of the product's halal status. Further research is the application of this application system to related institutions. Furthermore, this application is expected to help the public quickly see the halal certification status of a product.

References

Kim, Yeong Gug, and Eunju Woo. 2016. "Consumer Acceptance of a Quick Response (QR) Code for the Food Traceability System: Application of an Extended Technology Acceptance Model (TAM)." *Food Research International* 85:266–72. doi: 10.1016/j.foodres.2016.05.002.

Tarjan, Laslo, Ivana Šenk, Srdjan Tegeltija, Stevan Stankovski, and Gordana Ostojic. 2014. "A Readability Analysis for QR Code Application in a Traceability System." *Computers and Electronics in Agriculture* 109:1–11. doi: 10.1016/j.compag.2014.08.015.

Tiwari, Sumit. 2017. "An Introduction to QR Code Technology." Pp. 39–44 in *Proceedings - 2016 15th International Conference on Information Technology, ICIT 2016*. Vol. 1. Bhubaneswar, India: IEEE.

Zhu, Shanhong, and Pei Tang. 2015. "The Design and Implementation of Eggs' Traceability System Based on Mobile QR Code." *Advance Journal of Food Science and Technology* 7(2):99–101. doi: 10.19026/ajfst.7.1274.

Job Scheduling Performance Issues and Solutions of Big Data Applications in Apache Spark: A Review

Hasmila Amirah Omar[a]*, Shahnorbanun Sahran[b], Nor Samsiah Sani[c] and Azizi Abdullah[d]

[a,b,c,d]Faculty of Information Science and Technology, The National University of Malaysia, UKM Bangi, 43600, Malaysia
*Email: hasmilaomar@gmail.com

Abstract

Big data analytics has been an active area of research recently due to the industrial market. It has a significant impact on processing large datasets to extract hidden patterns and information for supporting decisions. Many big data processing frameworks have been designed to achieve the aims. Among the popular ones that are mostly adopted is the Apache Spark. It has become one of the widely used open-source processing frameworks for large-scale datasets. It is due to the framework capability that uses in-memory computations that enable fast processing. The key factor behind this efficient processing is the job scheduling management system. It is the crucial component in managing resources that influence the execution of any big data applications. Scheduling jobs is a challenging task, especially when the deployed cluster has different hardware capacities and incoming jobs can be heterogeneous. Some important factors need to be considered to further improving the performance of this framework. Hence, this paper aims to provide a review of Apache Spark in terms of scheduling performance challenges and previous efforts solutions. The analysis of this paper may provide better insights and a roadmap for further enhancement of job scheduling in Apache Spark.

Keywords: Big data analytics; Apache Spark; Job scheduling; Performance issues

1. Introduction

Big data is a term that refers to the large volume of data that grow exponentially with time (Khalil et al., 2020). The data can be in the form of structured, semi-structured and unstructured data collected by organisation to be analysed for finding informative insights that lead to better decisions. However, the massive and complex data exceeded the ability of traditional computing power to manage and capture the hidden potentials. Therefore, one of the top big data processing framework in use today such as Apache Spark, an open-source framework that can quickly process large-scale data sets in parallel (Zaharia et al., 2010). It utilised in-memory computing features that support batch, iterative, interactive and streaming applications, which are useful for complex computations to have different computation modes in one platform. An important feature of Spark is its RDD (Resilient Distributed Datasets). RDD is a distributed set of read-only elements, which can only be generated by deterministic operations.

Although Apache Spark is relatively good compared to other frameworks, it also has a performance bottleneck in job scheduling (Islam et al., 2020). It is difficult for beginners to comprehend the scheduling performance issues and the research solutions behind it instantly to further improve the performance. Therefore, this paper aims to provide a concise source of Apache Spark information, specifically on its scheduling performance bottleneck. We highlight the previous and some recent works about Apache Spark's enhancement based on the identified issues. Thus, providing some development directions for framework optimization. The rest of this paper is organized as follows. Section 2 describes Apache Spark's background, which includes its features and job processing flow. Section 3 presents reviews of Spark scheduling performance issues which leads to solutions from past researches. Finally, section 4 presents our conclusion of this review paper.

2. Job Scheduling Performance Issues and Solutions in Spark

The goal of job scheduling in Spark resource management is to plan the execution of tasks throughout the nodes. It aims to maximize resource utilization while minimizing the total execution time. This section will elaborate on the performance challenges or issues in Spark job scheduling and solutions available in the literature. We classify these issues into three categories, as shown in Table 1. All the categories can be described as follows:

2.1 Parameter Configuration

Parameter configuration refers to the setting of Spark parameter values before executing an application (Zaharia et al., 2010). Typically, the configuration of the Spark parameters can be done in 2 ways. One, the user can manually set the configuration, and two, they can use the default configuration to have an easy implementation. One of the issues found here is that it can cause slowdowns or even worst failures in the Spark applications if its parameters are not properly configured. Therefore, given the high significance of the problem, many previous efforts have been made to determine the optimal solutions for parameter configuration.

Among these, Petridis et al. (2017) applied manual tuning by trial and error to tune Spark configuration parameters. They conducted a series of experiments for all the possible combinations of parameters by utilizing expert knowledge to search for an optimal configuration. The results showed that their manually tuned method could increase Spark performance by 10 times speedup. Gounaris and Torres (2018), on the other hand provide an alternative approach to Petriditis et al. (2017), where they proposed a systematic methodology for parameter tuning. However, this study also involves with repeated experiments of a trial and error approach, but it is guided by a systematic methodology. The results of this study reveal that the proposed methodology improves the speed by up to 20 % during implementation compared to the default settings. However, both of these studies clearly is time-consuming as it needed considerable effort in performing repeated experiments to find the best parameter configurations. Furthermore, it requires expert knowledge and researcher experience to determine the value of the parameters at the beginning of the phase.

Study by Bian et al. (2014), proposed CSMethod, a simulator for Spark where the whole Spark application execution environment is simulated. This paper aims to provide a fast and accurate simulator as well as providing a reliable approach for testing parameter combinations until the optimal setting is met. This approach, however, seems rather challenging to precisely simulate the environment due to the vast hardware diversity and software complexity. Moreover, when applied to the actual cluster, there might be an inconsistency of getting the expected results due to a different implementation environment. Other techniques by Perez et al. (2018) developed a multi-parameter tuning method called PETS (Parameter Ensemble Table for Spark) using a Fuzzy approach. It utilized a metric called bottleneck score with multiple fuzzy engines and a parameter ensemble table. Most of the rules and fuzzy classes require knowledge from researchers or experts. PETS is able to tune 18 parameters simultaneously and outperformed other machine learning techniques with a speedup of up to x4.78 using 6 different workloads of the Hibench benchmark. However, there is a trade-off between performance speedup and convergence speed. Achieving a higher speedup resulted in slower convergence when compared to simple strategy due to high rates of changing the parameter at one time.

A more popular method uses machine learning-based approach by building models and making predictions on the performance before the application started. Previous efforts by Bao et al. (2019) proposed an automatic parameter tuning called Autotune. The researchers implement testbeds that use a sampling strategy called Latin Hypercube Sampling (LHS) to generate more samples based on given time constraints to train the model. Therefore, more promising configurations can be found using the trained prediction model. Autotune demonstrated that it improves execution time to 63.7% on average when compared to default parameter configuration. However, when compared to other tuning methods, the speedup improvement is only 6-24%. Other research that also utilized the LHS sampling strategy is Nguyen et al. (2018). Unlike in Autotune, they

applied the LHS technique to minimize the number of training samples and use a recursive random search algorithm to tune the parameter configurations. The results demonstrated that their proposed method reduces the execution time by 22.8% to 40% on 9 different applications compared to the default settings.

In Wang et al. (2017), the idea is to used binary and multi-class classification algorithms to predict the execution time under a given set of parameters. Data from actual executions for each workload is collected using random sampling to train the model. Their proposed method improved the time performance of an average of 36% lower running times when compared to the default settings. However, this technique needs to have intensive training for every specific workload to achieve the optimal model. A study by Gu et al. (2018) proposed tuning Spark parameter configurations for streaming applications using neural networks. They generated training data set randomly and used a random forest algorithm to build a prediction model to predict the execution time. On the other hand, the neural network approach is used to search for the optimal configuration based on the prediction model. The experimental results show that the proposed approach increases the performance of Spark streaming to 42.8% when compared with the default parameter configuration. Recent study by Li et al. (2020) proposed the ATCS system, an automated tuning approach using the Generative Adversarial Network (GAN) algorithm. The GAN algorithm is used to build a performance prediction model by reducing the model's complexity using less training data. They implemented a Random Parameter Generator (RPG) to produce random configurations for each workload as training data for the prediction model. The results show that Spark's performance can be improved by an average of 3.5 to 6.9 times compared to the performance of the default parameters.

Based on the previous works above, we can conclude that the experimental-based or trial and error approach is less effective due to high-dimensional parameter space and time consuming as it requires intensive repetition of experiments to test each combination of parameters. On the other hand, the simulation approach is a faster way to test all the parameters combination. However, it is challenging to simulate the real Spark environment as it requires depth knowledge of Spark internal systems to build one. Machine learning methods are gaining much popularity among researchers to facilitate better results in getting optimal configurations. More efficient machine learning methods should be explored to tackle this issue. Focusing not only on the improvement of the prediction accuracy and time performance, but it is also important to have estimation prediction of the usages of cores, memory, disk, and network before launching the application to ensure that all the resources are fully utilized.

2.2 Workload Characteristics

Workload characteristics refer to the job characteristics, i.e., the size of data, type of data (e.g. SQL or machine learning tasks, etc) or the resource requirements needed to run the data. Default Spark schedulers such as FIFO does not consider the workload characteristics in the scheduling decision to have an easy and straightforward implementation. By using this approach, it is hard to achieve efficient utilization of resources. In this section, we examine different methods and techniques to enhance Spark scheduling performance based on workload characteristics-aware.

Mao et al. (2019) proposed Decima that aims to improve the existing heuristic approach of task scheduling by considering the workload characteristics. It uses reinforcement learning (RL) and neural networks to learn scheduling policy through experience. It represents the scheduler as an agent that can learn from workload and cluster conditions without relying on incorrect assumptions. Decima encodes its scheduling strategy by observing the environment, taking action and improving its policy over time to make better decisions. The agent will be rewarded after taking any action, and the reward is set based on the scheduling objective (e.g., minimize average job completion time). The results show Decima improves average job completion time by 21% over default schedulers. However, the authors do not mention whether it supports for multi-tenancy framework, which is important for high-performance computing workload.

Liang et al. (2018) proposed a methodology of WSMC (workload-specific memory capacity) that overcame the problem of considering the workload characteristics based on memory capacity. Workload characteristics may have diverse memory capacity requirements. They use a metric of data expansion ratio as the input data for the workload classification. They also established the memory requirement prediction model for each workload and achieved the performance improvement of 40% compared to the manual configuration. This work, however, focusing on managing the memory space more efficiently rather than managing the executor, which is of importance to improve the time performance of the cluster. More recent work by Zaouk et al. (2021) used deep neural networks to develop a performance prediction model by embedding the workload characteristics. Given the diversity of the workload, they used collected run-time learning via representation learning. The goal of modeling is to derive a job-specific prediction model using the previous observations. They built an optimizer that automatically recommends configurations for the subsequent execution to improve the performance. Based on the result, they achieved a performance improvement of 52.4% on Spark streaming workloads compared to the baseline study.

To summarize, the user often underestimates the impact of the workload requirement in their scheduling decision. However, it is evidenced from previous works that it can cause low resource utilization. Although there are advanced methods of using deep neural networks and reinforcement learning to estimate the resources based on the workload, the performance needs to be improved. For future work, in-depth analysis and research are needed to explore on how a scheduler could consider dynamic workload where the resource requirement varies significantly during execution in multi-tenant environment.

2.3 Partition Size

Partition size refers to the level of parallelism in Spark schedulers. Scheduling in Spark allows running multiple tasks in parallel across a cluster of machines. By default, the number of partition is given based on the number of HDFS blocks of that file. If the partition size is too few, the scheduler cannot utilize all the cores available in the cluster. In contrast, there will be excessive overhead in managing small tasks. The optimal solution is to find a reasonable size partition that can utilize all the cores available and avoid overhead. In this section, we summarize the solutions gathered from the literature.

Study by Hernandez et al. (2018), provide solutions to find an optimal partition size configuration by using machine learning methods. The aim is to optimize the task parallelism of the application by predicting the optimal number of tasks per executors and tasks per machine. The authors used regression methods and based on the result, they achieved a 51% gain in performance when using the recommended settings. However, the approach here does not consider different machine specifications in a cluster known as a heterogeneous environment, a common environment in the distributed processing system.

Work by Wang et al. (2018), proposed a speculative mechanism to achieve optimal parallelism for Spark scheduler using a simulator known as STLS (Software Thread-Level Speculation) using the simulated annealing algorithm. The idea is to partition the input data in a fine-grained way and assign a number of threads in the cluster with small-scale data. Based on the results, they have achieved a 15-23% speedup compared to the baseline methods. However, this approach may be time-consuming as the design and implementation of Spark's execution model is far more complex. In Wang et al. (2019), the authors proposed a performance model to estimate the application's execution time for a given partition size. This paper aims to find an optimal partition size that can reduce the application execution time. The idea is to predict the possible straggler tasks or skewed task distribution by running with a fraction of input data. If the model predicts straggler tasks, the performance model will repartition the input data by adjusting the partition size. On the other hand, if there are skewed tasks, it will tune the locality setting. Based on the results, this paper demonstrates a performance improvement of 71%. However, this method can be costly when repartitioning a large amount of data as it will need to reshuffle the data to ensure it is balanced across the partitions. Gounaris et al. (2017) proposed a novel algorithm for configuring dynamic partitioning using Greedy and randomized approaches. The idea is to modify the degree

of parallelism or partition size during execution. They performed profiling activity to describe the application behavior as a function of the number of machines used in order to derive the dynamic partition solutions. Based on the results, the time performance improved to 50% using the estimated dynamic partition. However, to obtain accurate profiles with only a few test runs is a challenging task.

In summary, suboptimal partitioning can cause resource wastage. Determining the right size is crucial to the scheduler as it will reduce the incoming overhead. Although repartitioning can be done to solve the bottleneck, the process can be costly as it will involve reshuffling the data. It is something that needs to be avoided as data will continue to rise at an unprecedented level. The natural way to solve this is by using sampling data or application profiling. However, this can be particularly challenging tasks due to inaccurate sampling results and profiling. Thus, this constitutes a direction for future work on how to achieve the best solution in partitioning.

Table 1. Summary of scheduling performance issues as well as the solutions available in the literature

Author (Year)	Issue	Solution Approach	Methodology	Limitation
Li et al. (2020)	Parameter Configuration	Machine learning	Applied GAN algorithm to reduce complexity by using less training data and inplement RPG to produce random configurations	Model accuracy need to be improved
Bao et al.(2019)	Parameter Configuration	Machine learning	Constructed testbeds that used sampling strategy (LHS) to generate more samples to train the model.	The speedup improvement is only 6-24% when compared to other tuning methods
Gu et al. (2018)	Parameter Configuration	Machine learning	Implement Neural Network to predict changes in parameter configurations	Only support single job to optimise at one time
Nguyen et al. (2018)	Parameter Configuration	Machine learning	Applied the LHS technique to minimize the number of training samples and use recursive random search algorithm	Need to generate more samples of training data to achieve the optimal setting
Gounaris and Torres (2018)	Parameter Configuration	Experiment-based	Conducted repeated experiments guided by a systematic methodology.	Time consuming and requires expert knowledge
Perez et al. (2018)	Parameter Configuration	Fuzzy	Utilized a metric called bottleneck score with multiple fuzzy engines and a parameter ensembel table	Slower convergence rate
Petriditis et al. (2017)	Parameter Configuration	Trial and error	Conducted a series of experiments for all the possible combinations of parameters	Time consuming and requires expert knowledge
Wang et al. (2017)	Parameter Configuration	Machine learning	Binary classification and multi-classification	Requires intensive training for every specific workload
Bian et al. (2014)	Parameter Configuration	Simulation	Created a simulator for Spark environment to test various parameter configuration	Challenging to simulate the real environment
Zaouk et al. (2021)	Workload characteristic	Machine learning	Used deep neural networks to develop performance prediction model by embedding the workload characteristics	The optimizer's recommendation is too optimistic due to extrapolation in a sparse search space.
Mao et al. (2019)	Workload characteristic	Machine learning	Used reinforcement learning (RL) and neural networks to learn workload-specific scheduling algorithms	Does not support multi-tenancy framework
Liang et al. (2018)	Workload characteristic	WSMC	Use metric of data expansion ratio to the input data for the workload classification	Focus on managing the memory space more efficiently, rather than managing the executor
Wang et al. (2019)	Partition size	Machine learning	Predicting the optimal number of tasks per executors and tasks per machines	Repartitioning data can be very expensive as it requires to reshuffled the data
Hernandez et al. (2018)	Partition size	Boosted Regression Tree	Predict the possible straggler tasks distribution by running with a fraction of input data.	Does not support heterogeneous machines

Table 1. (Continued)

Author (Year)	Issue	Solution Approach	Methodology	Limitation
Wang et al. (2018)	Partition size	Simulated annealing	Partition the input data in a fine-grained way and assign number of threads in the cluster with small scale data	Time-consuming as the design and implementation far more complex
Gounaris et al. (2017)	Partition size	Greedy and Randomized	Performed profiling to modify the partition size during execution	Challenging tasks to obtain an accurate profiles with only few test runs

3. Conclusion

Job scheduling is the most crucial element in any data processing framework. It plays a vital role in achieving efficient utilization of resources. Existing scheduling solutions need to keep evolving as to properly support new challenges that keep arising. These necessities are important to facilitate a higher performance of data-intensive workload. In this paper, we present a review of Spark scheduling performance issues and compile the solutions available in the literature accordingly. We provide an analysis of the related works to date and suggestions for research directions. We hope that our effort provides an entry point for researchers to build a roadmap for future work to improve Spark scheduling performance.

Acknowledgements

This work is supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/UKM/02/6).

References

Bao, L., Liu, X., & Chen, W. (2019). Learning-based Automatic Parameter Tuning for Big Data Analytics Frameworks. In Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018. https://doi.org/10.1109/BigData.2018.8622018

Bian, Z., Wang, K., Wang, Z., Munce, G., Cremer, I., Zhou, W., … Xu, G. (2014). Simulating big data clusters for system planning, evaluation, and optimization. In Proceedings of the International Conference on Parallel Processing. https://doi.org/10.1109/ICPP.2014.48

Gounaris, A., Kougka, G., Tous, R., Montes, C. T., & Torres, J. (2017). Dynamic configuration of partitioning in spark applications. IEEE Transactions on Parallel and Distributed Systems. https://doi.org/10.1109/TPDS.2017.2647939

Gounaris, A., & Torres, J. (2018). A Methodology for Spark Parameter Tuning. Big Data Research. https://doi.org/10.1016/j.bdr.2017.05.001

Gu, J., Li, Y., Tang, H., & Wu, Z. (2018). Auto-Tuning Spark Configurations Based on Neural Network. In IEEE International Conference on Communications. https://doi.org/10.1109/ICC.2018.8422658

Hernández, Á. B., Perez, M. S., Gupta, S., & Muntés-Mulero, V. (2018). Using machine learning to optimize parallelism in big data applications. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2017.07.003

Islam, M. T., Srirama, S. N., Karunasekera, S., & Buyya, R. (2020). Cost-efficient dynamic scheduling of big data applications in apache spark on cloud. Journal of Systems and Software, 162, 110515. https://doi.org/10.1016/j.jss.2019.110515

Khalil, W. A., Torkey, H., & Attiya, G. (2020). Survey of Apache spark optimized job scheduling in big data. International Journal of Industry and Sustainable Development (IJISD) (Vol. 1). Retrieved from http://ijisd.journals.ekb.eg39

Li, M., Liu, Z., Shi, X., & Jin, H. (2020). ATCS: Auto-Tuning Configurations of Big Data Frameworks Based

on Generative Adversarial Nets. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2979812
Liang, Y., Chang, S., & Su, C. (2018). A workload-specific memory capacity configuration approach for in-memory data analytic platforms. In Proceedings - 15th IEEE International Symposium on Parallel and Distributed Processing with Applications and 16th IEEE International Conference on Ubiquitous Computing and Communications, ISPA/IUCC 2017. https://doi.org/10.1109/ISPA/IUCC.2017.00080
Mao, H., Schwarzkopf, M., Venkatakrishnan, S. B., Meng, Z., & Alizadeh, M. (2019). Learning scheduling algorithms for data processing clusters. In SIGCOMM 2019 - Proceedings of the 2019 Conference of the ACM Special Interest Group on Data Communication. https://doi.org/10.1145/3341302.3342080
Nguyen, N., Maifi Hasan Khan, M., & Wang, K. (2018). Towards Automatic Tuning of Apache Spark Configuration. In IEEE International Conference on Cloud Computing, CLOUD. https://doi.org/10.1109/CLOUD.2018.00059
Perez, T. B. G., Chen, W., Ji, R., Liu, L., & Zhou, X. (2018). PETS: Bottleneck-aware spark tuning with parameter ensembles. In Proceedings - International Conference on Computer Communications and Networks, ICCCN. https://doi.org/10.1109/ICCCN.2018.8487324
Petridis, P., Gounaris, A., & Torres, J. (2017). Spark parameter tuning via trial-and-error. In Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-319-47898-2_24
Wang, G., Xu, J., & He, B. (2017). A Novel Method for Tuning Configuration Parameters of Spark Based on Machine Learning. In Proceedings - 18th IEEE International Conference on High Performance Computing and Communications, 14th IEEE International Conference on Smart City and 2nd IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2016 (pp. 586–593). https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0088
Wang, K., Khan, M. M. H., Nguyen, N., & Gokhale, S. (2019). A Model Driven Approach Towards Improving the Performance of Apache Spark Applications. In Proceedings - 2019 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS 2019. https://doi.org/10.1109/ISPASS.2019.00036
Wang, Z., Zhao, Y., Liu, Y., & Lv, C. (2018). A speculative parallel simulated annealing algorithm based on Apache Spark. Concurrency Computation . https://doi.org/10.1002/cpe.4429
Zaharia, M., Chowdury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing. USENIX Association. https://doi.org/10.1017/CBO9781107415324.004
Zaouk, K., Song, F., Lyu, C., & Diao, Y. (2021). Neural-based Modeling for Performance Tuning of Spark Data Analytics. Retrieved from http://arxiv.org/abs/2101.08167

Explainable Recommender – Implementation Approaches

Neeraj Tiwary[a*], Shahrul Azman Mohd Noah[a], Fariza Fauzi[a], Steffen Staab[b]

[a]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
[b]WAIS, University of Southampton, UK & Analytic Computing, Universität Stuttgart, DE
* Email: p105920@siswa.ukm.edu.my, Neeraj.tiwary@gmail.com

Abstract

During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other web services, recommender systems are playing a more important role in our lives and are well-researched. Organizations want to recommend the contents/products based on individual needs. Personalization is the one aspect where every organization is paying more attention. Personalized services enable users to find related/relevant information quickly, be it a shopping item, movie, news, travel ideas, or restaurants that best suits their tastes from the countless choices. Thus, the demand for personalized recommenders has become pervasive in all domains. Besides providing the right recommendations, explainability for those recommendations is also very much required. The ability to explain how/why these recommendations are suggested will bridge the confidence and trust of the users in the system. Knowledge graph, reinforcement learning, and language models are some recent state-of-the-art technological advancements. Recommendation systems can benefit from these advancements by effectively applying them to enhance the efficacy of the explainable recommendations. The objective of this paper is to discuss the various implementation methods for developing an explainable recommendation engine. The contribution of this paper is to explain the concept of explainable recommendation as well as provide different implementation mechanisms and niche advancements in the explainable recommendation area.

Keywords: Explainable AI; Knowledge Graph; Reinforcement Learning; Language Models

1. Introduction

Recommender systems (RS) are algorithms aimed at suggesting relevant items (i.e. movies, books, products, etc.) to users based on his/her preferences towards those items (Cosley et al., 2003). They are critical in many industries like entertainment, e-commerce, media, and advertisements and pivotal in driving huge revenue for organizations. Some of the initial approaches for developing recommender systems were using content-based or collaborative filtering-based algorithms. Content-based recommendation is in which a user is recommended items that are similar to those that the user liked in the past, whereas collaborative recommendation is where a user is recommended items that other users with similar tastes liked in the past (Ricci et al., 2011). One of the main issues that these recommender systems has is the explainability of the recommendations. Though the concept of explainable recommendation was formally introduced by Zhang et al. 2014, this concept is not new, and many other researchers (Schafer, 1999; Herlocker, 2000) coined this idea earlier. Schafer et al., 1999 explained the recommendations reasoning through the customer past relationships with the e-commerce system. Herlocker et al., 2000 conducted a study on collaborative filtering algorithms in the MovieLens dataset and analyzed the explainability based on user surveys.

Recently, a series of AI regulations have entered into force, such as the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act of 2018, which emphasize the “right to explanation” of algorithmic decisions. Overall, the explainability of AI systems is always an imperative topic. This paper is organized to explain past and recent implementation approaches for explainable recommendation systems and niche advancements in this area.

2. Implementation Approaches

2.1. Early approaches

Content-based (CB) or collaborative filtering (CF) based recommendation algorithms are pioneer algorithms in the early approaches to personalize recommender systems (Ricci et al., 2011).

2.1.1. Content-based

Content-based algorithms use various available content information such as the utility, price, color, brand, size, etc. of the goods in e-commerce (Schafer, 1999), or the genre, director, duration, language, casting, etc. of the movies in review systems (Balabanovic, 1997) to model recommender systems. It is intuitive to explain the recommendation to users through various available item content information. Ferwerda, 2012 provided a complete study for elucidation of CB recommendations. CB algorithms have the main demerit of collecting various content information in different application domains, and it is a time-consuming effort.

2.1.2. Collaborative Filtering-based

Collaborative filtering-based approaches (Ekstrand, 2011) have addressed the issue associated with the CB approach. User-based CF (UBCF) represents each user as a vector of ratings and predicts the missing rating based on the weighted average of other users' ratings (Resnick, 1994). Item-based CF (IBCF), represents each item as a vector of ratings and predicts the missing rating based on the weighted average ratings from similar items (Sarwar, 2001). The explainability of CF lies in its design. UBCF can explain the behavior as "users who are similar to you loved this item", and IBCF can explain as "the item is similar to your previously loved items". CF approaches have improved the accuracy of the predictions but have less explainability (Herlocker, 2000a) and an issue of cold start (J. Gope, 2017), where it can't recommend the contents to new users.

2.2. Advanced approaches

The RS have been further enhanced through advanced techniques like latent factors (LFM), deep learning (DL), knowledge graph (KG), reinforcement learning (RL), and language models (LM) (Shaoxiong, 2021).

2.2.1. Latent Factor Models

The CF approach discussed in the previous "Early approaches" section has been further enhanced through dimensionality reduction methods like latent factor models (LFM). The most well-known LFM is matrix factorization (MF). MF methods have many specific techniques like singular value decomposition (Koren, 2009), non-negative MF (DD Lee, 2001), and probabilistic MF (A Mnih, 2008). These approaches create a latent factor representation to calculate the matching score of the user-item pairs. LFM especially MF and its variants were very successful in rating prediction tasks. Latent factors in LFMs do not possess any intuitive meanings, which restrict the explanatory power of recommendations. It necessitates the need for explainable recommendations (Zhang, 2018) efforts. Zhang et al., 2014, defined the explainable recommendation problem by aligning the latent dimensions with explicit features and proposed an Explicit Factor Model.

2.2.2. Deep Learning Models

CF methods have been further improved by the usage of deep learning (DL). Similarity learning and representation learning are the two main approaches in DL methods of CF approaches. In similarity learning, it

creates user-item embeddings and computes user-item similarity matching scores (X. He, 2017). The representation learning approach learns much richer user-item representations (Wu. L, 2019). The advantages of using DL for representation learning are twofold: (1) it reduces the efforts in handcraft feature design; and (2) enables recommendation models to include heterogeneous content information such as text, images, audio, and video. Deep representation learning helps to improve the search and recommendation performance but lacks transparency and explainability. The recommendations are hardly explainable to system designers/users.

2.2.3. Knowledge Graph

After LFM, the next big effort in the recommendation engine is the applicability of the KG. The KG improves prediction accuracy and enhances the explainability of the models. KG embeddings (KGE) have started leading the effort in explainable recommendation engines. Few researchers start focusing on TransE (Bordes, 2013) and node2vec (Grover, 2016) models. The objective of KGE is to find the similarity between entities by calculating their representation distance (Zhang F., 2016). The main issue with KGE models is that they produce black-box recommendations. Ai et al. 2018 enhanced the CF approach over KGE and later adopted a soft matching algorithm to find explanation paths between users and items. Here the explanation is not produced according to the reasoning process, but it is a post-hoc explanation. It is computationally expensive to traverse all the paths between a user-item node pair for similarity calculation in real world KG.

2.2.4. Reinforcement Learning

RL is another big area, and many successful applications belong to this area e.g. AlphaGo (D Silver, 2016), self-driving cars (Chopra, 2020) etc. It works under the framework of State, Agent, and Actions. In every state, the agent takes an appropriate action to maximize the rewards and go to another state. This framework belongs to an environment and demonstrates the ability of the agent to understand the high-level causal relationships. Xian et al., 2019 proposed a RL model for pathfinding to address the KG traversal issue for explainable recommendations. In the training stage, the agent finds the right user-item path with high rewards. In the inference stage, the agent will use the high reward path for recommendations. It thus addresses the issue of enumerating all the paths between user-item pairs. Using RL for recommendation is a fairly novel field.

2.2.5. Language Models

Large-scale pre-trained language models like OpenAI GPT and Bidirectional Encoder Representations from Transformer (BERT) have achieved great performance on a variety of language tasks using generic model architectures. Language models like BERT, GPT, ELMO, and many others are the main drivers for having semantic implementations and have a great scope in the recommendation engine (Shaoxiong, 2021). (Fei Sun, 2019) explained the sequential recommendation by using BERT. The research of generating natural language (NLG) explanation is still in its early stage, and LM have a great scope in enhancing the same.

3. Conclusion

Explainable recommendation, which provides explanations about the recommendations made to the user, has attracted increasing attention due to its ability in helping users make better decisions and increasing users' trust in the system. The idea of this paper is to understand various implementation approaches, their merits, and demerits for the successful implementation of an explainable recommendation engine.

Content and CF algorithms are the early approaches for developing recommendation engines. CB algorithms define user preferences based on the content information and the user's past interaction with the system. The model can capture the user-specific interests and recommend niche items. The main demerit with this algorithm is the domain-specific nature of the recommendation. CF algorithms have addressed this domain-specific issue. This algorithm is mainly using the wisdom of the crowd. The main demerits associated with CF algorithms are the cold-start problem and lesser explainability power than CB algorithms.

With the advancements in technologies, newer methods like latent factors, MF, and DL have enhanced the overall recommendation accuracy. The problem with these models lies in its transparency and explainability. The other known issues associated with these algorithms are cold-start and data sparsity. Various state-of-the-art technology advancements have been presented in the last few years. The KG can address cold-start, data sparsity, and explainability issues associated with the previous methods. RL may assist in identifying the right path over the KG for providing better recommendations to the end-user. Language models may assist in building a better contextual KG to enhance the overall accuracy of the user recommendations.

To conclude, these niche technologies have a great scope and will form a base for future research in the explainable recommendation area. The research in KG and RL are still in the nascent stages. LM seem to have a great scope in the recommendation engine and possess a great interest in the research community.

Acknowledgements

This work is supported by the UKM through the Prime Impact Fund under Grant DIP-2020-017.

References

Ai Q, et al. 2018. “Learning Heterogeneous Knowledge Base Embeddings” In: Algorithms, v:11, p:137 (link).
Andriy Mnih, et al. 2008. “Probabilistic matrix factorization”. In: NIPS ACM. 1257–1264.
Balabanovic, et al. 1997. “Fab: content-based, collaborative recommendation”. In: ACM. 40(3): 66–72.
Bordes Antoine, et al. 2013. “Translating embeddings for multi-relational data”. In NIPS 2787–2795.
Chopra R., et al. 2020 “End-to-End Reinforcement Learning for Self-driving Car”. In: ACIE vol 1082 (link)
D. Cosley, et al., 2003 “Is seeing believing? How recommender system interfaces”. In: SIGCHI, 585–592.
Daniel D Lee, et al. 2001. “Algorithms for non-negative matrix factorization”. In: NIPS ACM 556–562.
David Silver, et al. 2016. “Mastering the game of Go with deep neural networks”. In: Nature 529, 484-489.
Ekstrand, et al. 2011. “Collaborative filtering recommender systems”. ACM 4(2): 81–173.
Fei Sun, et al. 2019. “BERT4Rec: Sequential Recommendation with BERT”. In 28 ACM 11 pages.
Ferwerda, B., et al. 2012. “Explaining Content-Based Recommendations”. regular paper.
Grover Aditya, et al. 2016. “node2vec: Scalable feature learning for networks”. In 22 ACM, 855–864.
Herlocker, J. L., et al., 2000 “Explaining collaborative filtering recommendations”. In: 2000 ACM. 241–250.
Herlocker, J. L., et al., 2000a. “Understanding and improving auto CF systems”. University of Minnesota.
J Ben Schafer, et al., 1999 “Recommender systems in e-commerce”. In: EC’99 1st ACM. Pp 158-166 (link).
Jyotirmoy Gope, et al. 2017. “A survey on solving cold start problem in RS”. In: ICCCA.(link)
Koren, Y., et al. 2009. “Matrix factorization techniques for recommender systems”. In: Computer 42-8 30-37.
Resnick, P., et al. 1994. “GroupLens: an open architecture for collaborative filtering”. In: ACM. 175–186.
Ricci, et al., 2011 “Introduction to recommender systems handbook”. In: RS handbook. Springer. 1–35.
Sarwar, B., et al. 2001. “Item-based collaborative filtering recommendation”. In: 10th ACM. 285–295.
Shaoxiong Ji, et al. 2021. “A Survey on Knowledge Graphs: Representation,” In: IEEE PP 99 (link)
Wu L., et al. 2019. “A context-aware user-item representation learning”. ACM (TOIS). 37(2): 22.
Xian Y., Z. Fu, et al. 2019. “Reinforcement Knowledge Graph Reasoning for Explainable”. In: 42 ACM.
Xiangnan He, et al. 2017. Neural Collaborative Filtering. In WWW’17. Pp 173–182.
Zhang Fuzheng, et al. 2016 “Collaborative knowledge base embedding for RS”. In 22 ACM, 353–362.
Zhang Y., et al., 2018. “Explainable Recommendation: A Survey and New Perspectives”In:arXiv:1804.11192
Zhang Y., et al., 2014. “Explicit factor models for explainable recommendation”. In: 37th ACM. 83–92.

Length-Controlled Abstractive Summarization Based on Summary Output Area Using Transfer Learning

Sunusi Yusuf Yahaya[a*], Nazlia Omar[b], Lailatul Qadri Zakaria[c]

[a,b,c] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia
* Email: yusufsunusi63@gmail.com

Abstract

The recent state-of-the-art abstractive summarization models based on encoder-decoder models generate precisely one summary per source text. Length controlled summarization is an important aspect for practical applications such as newspaper or magazine cover slots summary. Some studies on length-controllable abstractive summarization use length embeddings in the decoder module for controlling the summary length while others use a word-level extractive module in the encoder-decoder model. Despite the fact that the length embeddings can control where to stop decoding, they fail to determine which information should be included in the summary within the length constraint. Providing a specific summary length can be helpful but not in some cases where the requirement is to fit the summary in a specific slot/area. Contrary to previous models, this paper aims to propose a length-controllable abstractive summarization model that incorporates an image processing phase which determines the area of the summary output slot to generate abstractive summary. The proposed model uses T5 transfer learning model to generate summary that perfectly fits the slots. The proposed model generates a summary in three steps. First, it uses opencv to determine the area of the given output slot where the summary will be displayed, for example in a newspaper cover slot. Secondly, the area is used to obtain the minimum and maximum length of the summary and; these will be used in T5 model to generate an abstractive summary that fits the summary output slot perfectly. Finally, self-attention mechanism was incorporated in the model to enhance the quality of the length controlled abstractive summary generated. Experiments with the CNN/Daily Mail dataset show that the proposed model is able to successfully perform the length-controlled summarization based on the computed summary output area.

Keywords: Natural Language Processing; Abstractive Text Summarization; Computer Vision; Summary Length Control.

1. Introduction

In recent years, there has been a great demand for the use of data obtained from a variety of sources including scientific literature, medical reports, and social networks. Text summarization is the process of generating a brief fluid summary of a longer text document. Constraining summary length, while largely neglected in the past, is actually an important aspect of abstractive summarization. For example, given the same input document, if the summary is to be displayed on mobile devices, or within a fixed area of advertisement slot on a website, editors may want to produce a much shorter summary. Unfortunately, most existing abstractive summarization models are not trained to react to summary length constraints (Yizhu et al., 2018).

Fan et al. (2017), who applies convolutional sequence to sequence model on multi- sentence summarization, converts length range as some special markers which are predefined and fixed. Unfortunately, this approach cannot generate summaries of arbitrary lengths. It only generates summaries in predefined ranges of length, thus only meets the length constraints approximately. Miao and Blunsom (2016) extended the seq2seq framework and proposed a generative model to capture the latent summary information, but they did not consider the recurrent dependencies in their generative model leading to limited representation ability. Magazine or newspaper editors tend to require summary that will fit into a certain slot in a cover, the current state-of-the-art as discussed above do not address output area based summary. For example, given a long newspaper story that need to be summarized to fit a portion in the newspaper cover, the previous work will not be able to provide summary for this since they are all based on specified number of summary word length.

Although length embeddings can control where to stop decoding, they do not decide which information should be included in the summary within the length constraint (Saito et al., 2020). However, length embeddings only add length information on the decoder side. Consequently, they may miss important information because it is difficult to take into account which content should be included in the summary for certain length constraints. This research will focus on developing an arbitrary length controllable abstractive text summarization model which will enable automatic text summarizations based on desired length constraint or output area constraint.

2. Related Work

Kikuchi et al. (2016) were the first to propose length embedding for length-controlled abstractive summarization. Fan et al. (2018) also used length embeddings at the beginning of the decoder module for length control. They present a neural summarization model with a simple but effective mechanism to enable users to specify these high level attributes in order to control the shape of the final summaries to better suit their needs. Liu, et al., (2018) proposed a CNN-based length-controllable summarization model that uses the desired length as an input to the initial state of the decoder. Takase and Okazaki (2019) introduced positional encoding that represents the remaining length at each decoder step of the Transformer-based encoder-decoder model. Saito et al. (2020) used extractive-and-abstractive summarization which incorporates an extractive model in an abstractive encoder-decoder model.

2.1. Critical Analysis of Previous Work on Length Control

As can be seen in Table 2.1, some of the previous works analyzed have achieved summary length control either pre-defined (Fan et al 2017; Kikuchi et al., 2016; Yizhu et al., 2018) or arbitrary (Takase & Okazaki, 2019; Makino et al., 2019; Saito et al., 2020). Despite having enhanced length constrained summarization quality, all the models require a specific length to be provided before summary is generated. In Saito et al., (2020), specific length of the prototype text must be given before it is inputted to their encoder decoder model for summary generation. Likewise in Takase and Okazaki (2019), remaining length must be defined at each decoder step of the Transformer-based encoder-decoder model.

Table 1: Literature on abstractive text summarization

Author	Title	Technique	Length Control	Area
Fan et al., 2017	Controllable Abstractive Summarization	Convolutional Seq2Seq Model	Yes	No
Yizhu et al., 2018	Controlling Length in AS Using a CNN	CNN Seq2Seq Model	Yes	No
Yao et al., 2019	Multi-Task Learning Framework for AS	Long Short-Term Memory (LSTM)	No	No
Yong et al., 2019	AS with a Convolutional Seq2Seq	Convolutional Seq2Seq Model	No	No
Petr et al., 2019	AS: A Low Resource Challenge	Transformer Model	No	No
Makino et al., 2019	Global Optimization under Length Constraint for Neural Text Summarization	CNN based encoder decoders	Yes	No
Takase & Okazaki, 2019	Positional Encoding to Control Output Sequence Length	neural encoder-decoder model, Transformer	Yes	No
Saito et al., 2020	Length-controllable AS by Guiding with Summary Prototype	Pointer-Generator, Prototype Extraction	Yes	No

In Figure 2.1 for example, given a long newspaper story that need to be summarized to fit a portion in the newspaper cover, the previous work will not be able to provide summary for this since they are all based on specified number of summary word length.

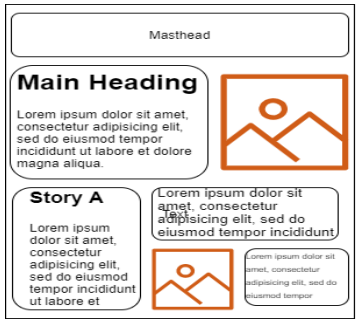


Fig. 1. Newspaper summary slots example

The downside of having predefined or arbitrary length is there will be cases where the usage of the summary is for a specific space/area. Magazine or newspaper editors tend to require summary that will fit into a certain slot in a cover. The current state-of-the-art as discussed above do not address output area based summary. Neither Saito et al., 2020 nor any of the other previous works are able to generate summary for the aforementioned scenario.

3. Methodology

Key features will be taken from the analysed techniques and will be used to develop an algorithm that will perform summarization based on length constraint. Finally, an intensive evaluation of the quality and relevance of the summary will be done in order to verify the enhancements. It is anticipated that by the end of the research, the new model will be more enhanced than the current length control techniques mentioned in this paper.

3.1. Output area based abstractive text summarization (Proposed)

Given the source text and an image with a portion where the summary will be portrayed (such as in magazine/newspaper story covers), using image processing the area of the summary output portion is calculated to determine the suitable summary length that will fit the portion. Next, the determined suitable length is embedded into the encoder-decoder model sequence to sequence technique to perform the summarization.

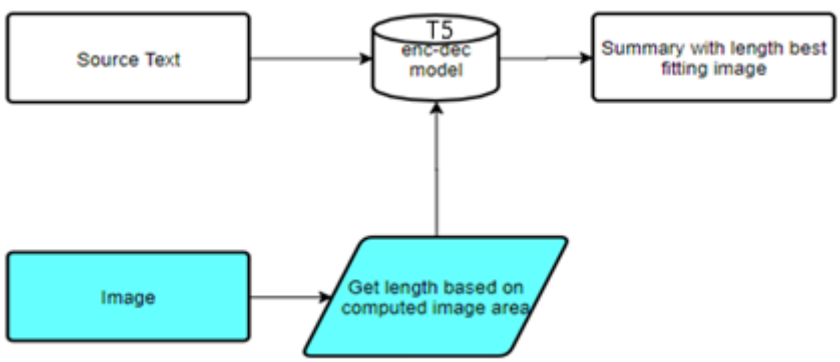


Fig. 2. Proposed model

The idea is to obtain a summary with a desirable length without having to predefine a specific length but rather produce the summary through analyses of where the summary will be outputted.

4. Results and Discussions

4.1 Computing Summary Output Area

Using the Opencv image processing library, the length and width of the summary output slot is obtain. The width and length of this portion is used to compute the area of the shape of the portion which will be used to determine the maximum length of the summary. In Figure 4.1, the maroon portion within the image is the designated slot where the summary will be displayed. Based on the image above the area of the portion is 0.36in which will be used as the maximum length of the output summary therefore making sure that summary fits the portion perfectly.

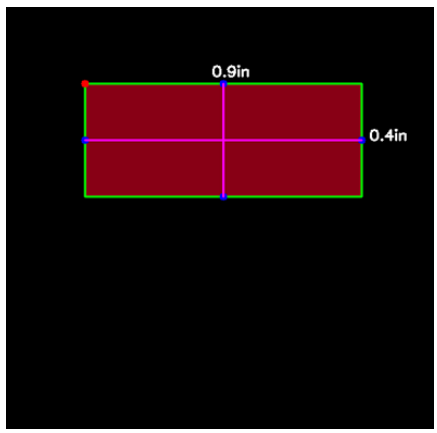


Fig. 3. Detected and analysed summary output area.

4.2. Generating Summary

The summary is generated by modifying the T5 (Text-To-Text Transfer Transformer) model and adding the area constraint. T5 is a new transformer model from Google that is trained in an end-to-end manner with text as input and modified text as output (Raffel et al., 2020). First, an input sequence of tokens is mapped to a sequence of embeddings, which is then passed into the encoder. The encoder consists of a stack of "blocks", each of which comprises two subcomponents: a self-attention layer followed by a small feed-forward network. Layer normalization (Ba et al., 2016) is applied to the input of each subcomponent. A simplified version of layer normalization is used where the activations are only rescaled and no additive bias is applied. After layer normalization, a residual skip connection adds each subcomponent's input to its output.

4.3. Length-controllable Summary

The computed area is used to obtain the minimum and maximum length of the summary and; these will be parsed in T5 model to generate an abstractive summary that fits the summary output slot perfectly. Using Figure 4.1, the area of 0.36in is parsed to the T5 model which generate then generates a summary that fits the specified output portion. The summary generated without output area constraint does not fit the designated slot of the summary. Meanwhile, the summary generated with the proposed area constraint model generates summary that fits desired portion.

5. Conclusion

The proposed model is expected to compute area of a specified portion within a given image. The proposed model will produce an abstractive summary using computed area from the image making it fit perfectly to the portion. The proposed model is expected to outperform other approaches and can be used for text summarization. Future works will include modifying other abstractive summarization techniques with the area constraint such as pointer generator, reinforcement learning and convolutional sequence to sequence models to determine which performs with greater results.

References

Fan, A., Grangier, D., & Auli, M. (2017). Controllable abstractive summarization. arXiv preprint arXiv:1711.05217.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., & Okumura, M. (2016). Controlling output length in neural encoder-decoders. arXiv preprint arXiv:1609.09552.

Li, P., Lam, W., Bing, L., & Wang, Z. (2017). Deep recurrent generative decoder for abstractive text summarization. arXiv preprint arXiv:1708.00625.

Liu, Y., Luo, Z., & Zhu, K. (2018). Controlling length in abstractive summarization using a convolutional neural network. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4110-4119).

Makino, T., Iwakura, T., Takamura, H., & Okumura, M. (2019, July). Global optimization under length constraint for neural text summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1039-1048).

Parida, S., & Motlicek, P. (2019, November). Abstract text summarization: A low resource challenge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5994-5998).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Rani, R. S., & Devi, P. L. (2020). A Literature Survey on Computer Vision Towards Data Science. 8(6), 3305–3309.

Saito, I., Nishida, K., Nishida, K., Otsuka, A., Asano, H., Tomita, J., Shindo, H., & Matsumoto, Y. (2020). Length-controllable Abstractive Summarization by Guiding with Summary Prototype. https://github.com/google-research/bert/

Schumann, R. (2018). Unsupervised abstractive sentence summarization using length controlled variational autoencoder. arXiv preprint arXiv:1809.05233.

Takase, S., Takase, S., & Okazaki, N. (2019). Positional encoding to control output sequence length. arXiv preprint arXiv:1904.07418.

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why self-attention? A targeted evaluation of neural machine translation architectures. ArXiv, 4263–4272.

Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. PressAcademia Procedia, 5(1), 205-213.

Yang, W., Tang, Z., & Tang, X. (2018, May). A hierarchical neural abstractive summarization with self-attention mechanism. In 2018 3rd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2018) (pp. 514-518). Atlantis Press.

Yolchuyeva, S., Németh, G., & Gyires-Toth, B. (2020). Self-attention networks for intent detection. ArXiv, 1373–1379.

Zhang, Y., Li, D., Wang, Y., Fang, Y., & Xiao, W. (2019). Abstract text summarization with a convolutional Seq2seq model. Applied Sciences, 9(8), 1665.

Text Encryption Based on DNA Cryptography, RNA, and Amino Acid

Omar Fitian Rashid

Department of Computer Technology Engineering, Al-Hikma University College, Baghdad, 10015, Iraq
* Email: omaralrawi08@yahoo.com

Abstract

To achieve safe security to transfer data from the sender to receiver, cryptography is one way that is used for such purposes. However, to increase the level of data security, DNA as a new term was introduced to cryptography. The DNA can be easily used to store and transfer the data, and it becomes an effective procedure for such aims and used to implement the computation. A new cryptography system is proposed, consisting of two phases: the encryption phase and the decryption phase. The encryption phase includes six steps, starting by converting plaintext to their equivalent ASCII values and converting them to binary values. After that, the binary values are converted to DNA characters and then converted to their equivalent complementary DNA sequences. These DNA sequences are converted to RNA sequences. Finally, the RNA sequences are converted to the amino acid, where this sequence is considered as ciphertext to be sent to the receiver. The decryption phase also includes six steps, which are the same encryption steps but in reverse order. It starts with converting amino acid to RNA sequences, then converting RNA sequences to DNA sequences and converting them to their equivalent complementary DNA. After that, DNA sequences are converted to binary values and to their equivalent ASCII values. The final step is converting ASCII values to alphabet characters that are considered plaintext. For evaluation purposes, six text files with different sizes have been used as a test material. Performance evaluation is calculated based on encryption time and decryption time. The achieved results are considered as good and fast, where the encryption and decryption times needed for a file with size of 1k are equal to 2.578 ms and 2.625 ms respectively, while the encryption and decryption times for a file with size of 20k are equal to 268.422 ms and 245.469 ms respectively.

Keywords: Cryptography; DNA cryptography; Encryption; Decryption; Security.

1. Introduction

With the modern applications and the increased use of the internet and network technology, the security threats are also increasing for the users, as accompanied by the large quantity of information transferred on the network. Various approaches have been used to break the system to steal important information. Therefore, security becomes a significant subject for those dealing with critical data and for modern computing systems. Recently, secret writing methods were used to save data from adversaries, and the most popular and most used method is cryptography (Kaundal and Verma, 2014). The computing technique of DNA cryptography can be used in encrypting or encoding the data to achieve safe transferring of the information. These capabilities are due to the DNA properties like parallel molecular computing, storing, transmitting the data, and computing capabilities (Bhimani, 2018). Cryptography was used to encrypt the data, and several kinds of research were proposed in this area. An asymmetric cryptography method based on a chaotic map and a multilayer machine learning network was also introduced (Lin at al., 2021). This system improved the security for real-time applications by randomly generating numbers to update secret keys using particular control parameters. A new mapping method was proposed by Keerthi and Surendiran (2017) to encode the message as points for the elliptic curve. The proposed technique begins by converting text to ASCII values, then converting these values to hexadecimal values. Then these values are used as x and y coordinates. A new method was introduced by Raghunandan et al. (2017) to remove the RSA algorithm drawback in terms of both integer fraction technique

and Wiener's attack that depending on decryption key calculation. This method differs from the RSA algorithm by the process of key generation, where this key is based on alpha, prime numbers, and Pell's equation. However, this led to making the gain of the private key from the public key is difficult. Viral et al. (2014) distributed a video to photo frames then stored it. Each frame includes three layers: red, blue, and green layers; each pixel is 8-bit values than pixel value may be any value with a range from 0 to 255. This system is performed by using two pixels for each frame to add text, and these pixels are the top left and the bottom right corner. After finishing all changes (the two pixels) for all frames, all images are ordered sequentially and converted to a video containing text encryption. Roy et al. (2017) proposed a novel text encrypting method by combining a Hopfield neural network and a DNA cryptographic model.

Firstly, a binary sequence is created and used for key generation, and then this sequence is created by using a chaotic neural network. The text is converted to ASCII values which in turn are converted to binary sequences. Then they encrypted them based on the transformation between chaotic neural network and permutation function. Singh and Singh (2015) introduced a new technique by removing the traditional choosing method for points in the elliptic curve. This system is done by converting text to their equivalent ASCII values, and these values are used as input for the Elliptic curve cryptography. The advantages of this system are reducing the cost and remove the common lookup table. UbaidurRahman et al. (2015) proposed a new DNA encoding idea, where the proposed method generated DNA sequences to encode text as DNA sequences. This text can include the alphabet, numbers, and some other characters. Abd El-Latif and Moussa (2019) proposed an encryption method that consists of two rounds and similar to the Data Encryption Standard. The method is used two keys to encrypt the message, the first key is induced from the elliptic curve cryptography, and the second key is achieved based on the mapping of second characters repeated in the key. Then, hide the ciphertext in the second DNA sequence. Raj and Panchami (2015) proposed a DNA-based cryptographic key generation method to produce keys for ciphering applications. This method is carried out by using a key with a size equal to the half size that is needed for a cryptographic key, and then four vectors are derived to represent the four DNA bases. Finally, the cryptographic key is created by using a linear formula for all DNA vectors. Kamaraj et al. (2016) proposed a double-layered security system, where this system is performed by encrypting the plaintext twice; the used key length is equal to 1000 characters. Then, they used DNA sequences to increase the cryptanalysis complexity.

The rest of the paper is organized as follows: Section 2 introduce our proposed cryptography method. The performance results of the proposed system and its discussion are shown in Section 3. Finally, Section 4 presented the conclusions of the current work.

2. Materials and Methods

The proposed cryptography method consists of two phases; encryption and decryption.

2.1. Encryption phase

Located at the sender side, and used to encrypt the message before sending it to the receiver, this phase contains sex steps of data conversion, as shown in Fig 1.

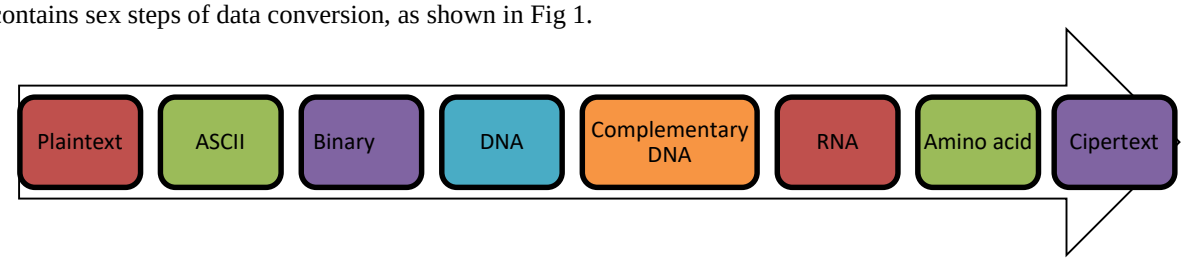


Fig. 1. Encryption steps applied in the encryption phase

Step -1: Convert each character in plaintext to their equivalent ASCII value (example: character A equals 65 in ASCII).

Step -2: Convert each ASCII value into binary format (example: 65 in ASCII equals 01000001 in binary).

Step -3: Convert every two binary values to one DNA character (either A, C, G, or T) based on table 1 (example: 00 in binary equals A as DNA character).

Table 1. Binary to DNA table

Binary	DNA
00	A
01	C
10	G
11	T

Step -4: Convert each DNA character to their complementary DNA character. Every A character is replaced with T and vice versa. Every C character is replaced with G and vice versa (example: the complementary DNA for character A equals T).

Step -5: Convert each DNA character to their equivalent RNA character. The only change is by converting a DNA character that equal to T to its equivalent RNA character that equal to U (example: the RNA for character T is equal to U).

Step -6: Convert each three RNA characters to amino acid value based on Table 2 (example: the amino acid for character A equals UUU or UUC). The amino acid table is built based on the amino acid idea shown in Fig 2 (Smith, 2008).

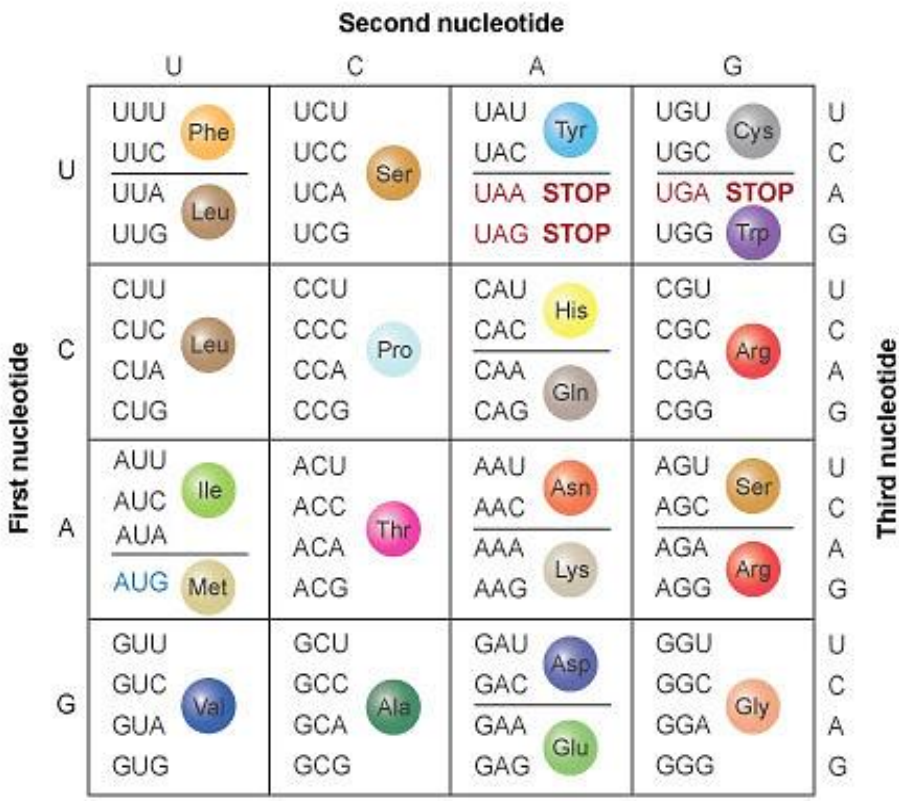


Fig. 2. Amino acids for RNA codons

Table 2. Amino acids and their corresponding RNA codons

Character	Protein sequences	Character	Protein sequences
A	UUU, UUC	N	CAU, CAC
B	UUA, UUG	O	CAA, CAG
C	CUU, CUC, CUA, CUG	P	AAU, AAC
D	AUU, AUC, AUA	Q	AAA, AAG
E	AUG	R	GAU, GAC
F	GUU, GUC, GUA, GUG	S	GAA, GAG
G	UCU, UCC, UCA, UCG	T	UGU, UGC
H	CCU, CCC, CCA, CCG	U	UGA
I	ACU, ACC, ACA, ACG	V	UGG
J	GCU, GCC, GCA, GCG	W	CGU, CGC, CGA, CGG
K	UAU, UAC	X	AGU, AGC
L	UAA	Y	AGA, AGG
M	UAG	Z	GGU, GGC, GGA, GGG

2.2 Decryption phase

Located at the receiver side, and it is used to decrypt a message received from the sender. This phase contains six steps which are the same steps on the sender side but in reverse order. These steps are:

Step -1: Convert each amino acid value to three RNA characters (example: the character A is equal to amino acid values either UUU or UUC).

Step -2: Convert each RNA character to their equivalent DNA character (example: the DNA for character U equals T).

Step -3: Convert each DNA character to their complementary DNA character (example: the complementary DNA for character A equals T).

Step -4: Convert each DNA character into two binary values (example: A equals 00).

Step -5: Convert every eight binary numbers to their equivalent ASCII value (example: 01000001 in binary equals 65 in ASCII).

Step -6: Convert each ASCII value to its equivalent character (example: 65 in ASCII is equal to character A).

3. Results and Discussions

In order to perform the above procedure of encryption to secure the transferred information, the following example will highlight the result of the currently proposed method.

Encryption steps example, where the plaintext is "**This is an example of encryption steps**".

Step-1: convert plaintext to ASCII: 84 - 104 - 105 - 115 - 32 - 105 - 115 - 32 - 97 - 110 - 32 - 101 - 120 - 97 - 109 - 112 - 108 - 101 - 32 - 111 - 102 - 32 - 101 - 110 - 99 - 114 - 121 - 112 - 116 - 105 - 111 - 110 - 32 - 115 - 116 - 101 - 112 – 115

Step-2: convert ASCII value to binary format: 01010100 - 01101000 - 01101001 - 01110011 - 00100000 - 01101001 - 01110011 - 00100000 - 01100001 - 01101110 - 00100000 - 01100101 - 01111000 - 01100001 - 01101101 - 01110000 - 01101100 - 01100101 - 00100000 - 01101111 - 01100110 - 00100000 - 01100101 - 01101110 - 01100011 - 01110010 - 01111001 - 01110000 - 01110100 - 01101001 - 01101111 - 01101110 - 00100000 - 01110011 - 01110100 - 01100101 - 01110000 – 01110011

Step-3: convert binary values to DNA sequences:
CCCACGGACGGCCTATAGAACGGCCTATAGAACGACCGTGAGAACGCCCTGACGACCGTCCTAA
CGTACGCCAGAACGTTCGCGAGAACGCCCGTGCGATCTAGCTGCCTAACTCACGGCCGTTCGTG
AGAACTATCTCACGCCCTAACTAT

Step-4: convert DNA to their complementary DNA:
GGGTGCCTGCCGGATATCTTGCCGGATATCTTGCTGGCACTCTTGCGGGACTGCTGGCAGGATTG
CATGCGGTCTTGCAAGCGCTCTTGCGGGCACGCTAGATCGACGGATTGAGTGCCGGCAAGCACT
CTTGATAGAGTGCGGGATTGATA

Step-5: convert DNA sequences to RNA sequences:
GGGUGCCUGCCGGAUAUCUUGCCGGAUAUCUUGCUGGCACUCUUGCGGGACUGCUGGCAGGA
UUGCAUGCGGUCUUGCAAGCGCUCUUGCGGGCACGCUAGAUCGACGGAUUGAGUGCCGGCAA
GCACUCUUGAUAGAGUGCGGGAUUGAUA

Step-6: convert RNA sequences to amino acid values, which is considered as the ciphertext for this example: "**ZTCHRDBHRDBCJCBWRTVORTEWGTQWGTZNJYGIRUFHJXICRYFWRU"**

Decryption steps example, where the ciphertext is
"ZTCHRDBHRDBCJCBWRTVORTEWGTQWGTZNJYGIRUFHJXICRYFWRU"

Step-1: convert amino acid values to RNA sequences:
GGGUGCCUGCCGGAUAUCUUGCCGGAUAUCUUGCUGGCACUCUUGCGGGACUGCUGGCAGGA
UUGCAUGCGGUCUUGCAAGCGCUCUUGCGGGCACGCUAGAUCGACGGAUUGAGUGCCGGCAA
GCACUCUUGAUAGAGUGCGGGAUUGAUA

Step-2: convert RNA sequences to DNA sequences:
GGGTGCCTGCCGGATATCTTGCCGGATATCTTGCTGGCACTCTTGCGGGACTGCTGGCAGGATTG
CATGCGGTCTTGCAAGCGCTCTTGCGGGCACGCTAGATCGACGGATTGAGTGCCGGCAAGCACT
CTTGATAGAGTGCGGGATTGATA

Step-3: convert DNA to their complementary DNA:
CCCACGGACGGCCTATAGAACGGCCTATAGAACGACCGTGAGAACGCCCTGACGACCGTCCTAA
CGTACGCCAGAACGTTCGCGAGAACGCCCGTGCGATCTAGCTGCCTAACTCACGGCCGTTCGTG
AGAACTATCTCACGCCCTAACTAT

Step-4: convert DNA sequences to binary values: 01010100 - 01101000 - 01101001 - 01110011 - 00100000 - 01101001 - 01110011 - 00100000 - 01100001 - 01101110 - 00100000 - 01100101 - 01111000 - 01100001 - 01101101 - 01110000 - 01101100 - 01100101 - 00100000 - 01101111 - 01100110 - 00100000 - 01100101 - 01101110 - 01100011 - 01110010 - 01111001 - 01110000 - 01110100 - 01101001 - 01101111 - 01101110 - 00100000 - 01110011 - 01110100 - 01100101 - 01110000 – 01110011

Step-5: convert binary values to ASCII value: 84 - 104 - 105 - 115 - 32 - 105 - 115 - 32 - 97 - 110 - 32 - 101 - 120 - 97 - 109 - 112 - 108 - 101 - 32 - 111 - 102 - 32 - 101 - 110 - 99 - 114 - 121 - 112 - 116 - 105 - 111 - 110 - 32 - 115 - 116 - 101 - 112 – 115

Step-6: Convert ASCII to alphabet characters, which is considered as the plaintext that was sent by the sender: "**This is an example of encryption steps**".

The proposed method consists of six levels of encryption, and this leads to keeps attackers from accessing the sending data, information, or messages, also make the process of break this system is more complicated. The performance evaluation for the proposed system is calculated based on encryption time and decryption time, where six text files are used with different sizes. These text files have sizes of 1K, 2K, 3K, 5K, 10K, and 20K. The obtained encryption time and decryption time for different file sizes are shown in Table 3 and Fig 3.

Table 3. Encryption and decryption times

Size	Total Plaintext Characters	Total Ciphertext Characters	Encryption Time (ms)	Decryption Time (ms)
1K	1001	1335	2.578	2.625
2K	1327	1769	3.406	3.734
3K	3000	4000	12.515	12.203
5K	5115	6820	25.24	31.094
10K	10229	13639	110.266	97.171
20K	20442	27256	268.422	245.469

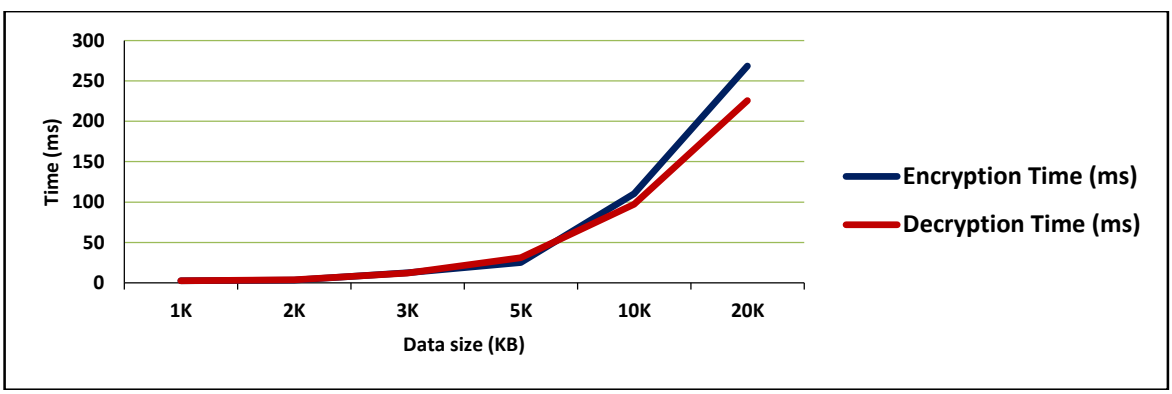


Fig. 3. Encryption and decryption times

4. Conclusion

A new cryptography system is proposed, this system is divided into two phases: encryption and decryption. The encryption phase has six steps, where firstly convert plaintext to their equivalent ASCII values, secondly convert ASCII values to binary values, thirdly convert binary values to DNA character, fourthly convert DNA sequences to their equivalent complementary DNA, fifthly convert these DNA sequences to RNA sequences, and finally convert RNA sequences to an amino acid, and this sequence sends it to the receiver as ciphertext. The decryption phase also includes six steps which are the same encryption steps but in reverse order. Six text files with different sizes have been used for evaluation purposes. Performance evaluation for the proposed system is calculated based on encryption time and decryption time. It concluded that the proposed method could be a successful one in this trend of increasing the security of data transfer.

References

Abd El-Latif, E. I., & Moussa, M. I., (2019). Information hiding using artificial DNA sequences based on Gaussian kernel function. Journal of Information and Optimization Sciences.

Bhimani, P. (2018). A Review on cryptography techniques using DNA computing. International Journal of Computer Engineering in Research Trends. 5(6), 187-191.

Kamaraj, A., Bhavithara, M., & Bhrintha, A. P. (2016). DNA-based encryption and decryption using FPGA. International Journal of Current Research and Modern Education (IJCRME'16), 89-94.

Kaundal, A. K., & Verma, A. K. (2014). DNA based cryptography: a review. International Journal of Information and Computation Technology, 4(7), 693-698.

Keerthi, K., & Surendiran, B. (2017). Elliptic curve cryptography for secured text encryption. 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 1-5.

Lin, C. H., Wu, J. X., Chen, P., Li, C. M., Pai, N. S., & Kuo, C. L. (2021) Symmetric cryptography with a chaotic map and a multilayer machine learning network for physiological signal infosecurity: case study in electrocardiogram. IEEE Access, 9, 26451-26467.

Raghunandan, K. R., Shetty, R., & Aithal, G. (2017). Key generation and security analysis of text cryptography using cubic power of Pell's equation. 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 1496-1500.

Raj, B. B., & Panchami, V., (2015). DNA-based cryptography using permutation and random key generation method. International Journal of Innovative Research in Science, Engineering and Technology, 3(5), 263-267.

Roy, S. S., Shahriyar, S. A., Asaf-Uddowla, M., Alam, K. M. R., & Morimoto, Y. (2017). A novel encryption model for text messages using delayed chaotic neural network and DNA cryptography. 2017 20th International Conference of Computer and Information Technology (ICCIT), 1-6.

Singh, L. D., & Singh, K. M. (2015). Implementation of text encryption using elliptic curve cryptography. Procedia Computer Science, 54, 73-82.

Smith, A. (2008). Nucleic acids to amino acids: DNA specifies protein. Nature Education, 1(1), 126.

UbaidurRahman, N. H., Balamurugan, C., & Mariappan, R. (2015). A novel string matrix data structure for DNA encoding algorithm. Procedia Computer Science, 46, 820-832.

Viral, G. M., Jain, D. K., & Ravin, S. (2014). A real-Time approach for secure text transmission using video cryptography. 2014 Fourth International Conference on Communication Systems and Network Technologies, 635-638.

Literature Review: Information Extraction Using Named-Entity Recognition with Machine Learning Approach

R Fenny Syafariani[a*], Rio Yunanto[b]

[ab]Universitas Komputer Indonesia, Jl. Dipatiukur No. 112-116, Bandung, Indonesia
Email:[a]r.fenny.syafariani@email.unikom.ac.id

Abstract

The purpose of this study is to help researchers identify and map machine learning algorithms from the results of previous studies with the theme of recognizing named-entities. This study's research method examines works of literature on the topic of introducing named-entities with the machine learning approach. The literature ranged from the year 2018 to 2020 and was collected through the use of Google Scholar. In this study, one of the critical research questions to be answered is whether machine learning algorithms have been used in named-entity recognition research. The introduction of named-entities is able to use three approaches: 1) machine learning, 2) deep learning, and 3) a combination of both. From the result, it was discovered that the combination of Conditional Random Field (CRF) machine learning and Bidirectional Long Short-Term Memory (Bi-LSTM) deep learning were used in 4 out of 7 analyzed works of literature.

Keywords: NER;information;extraction;named-entity;review;

1. Introduction

Entity extraction process is widely known to be one of the important stages in information extraction. As one of the methods, named-entity recognition can automatically extract entities in a particular text and determine its category. It includes extracting object name, object, person, or company name (Wibisono & Khodra, 2018). As an example, from the sentence "Flood and landslide in Nganjuk, 23 people reported missing", the recognition process will result in a named-entity (often referred to as a mention) with "Nganjuk" as the type of location as well as "Flood" and "landslide" as the type of event. It shows that the named-entity recognition process is able to automatically recognize entities in a sentence or text and is able to categorize the entity according to the type referred to in the text.

One of the ways that named-entity recognition can be done is through the formulation of a certain word or phrase patterns. For example, the typical word pattern of the phrase “come from…” or “go away from…” would be followed by location-type entity words. Various combinations of word patterns can be taught in machine learning using training data to build knowledge on the algorithms used. Therefore, it further supports the fact that the introduction of machine learning-based named-entity recognition will be able to detect named entities automatically (Giarsyani, 2020).

We have conducted a literature study to determine a suitable machine learning algorithm that could open up new research areas opportunities. Through literature study, we have created research questions as a guide in the research process which includes: 1) What objects or datasets have been used in the research on recognizing named-entity?, 2) What machine learning algorithms have been used in named-entity recognition research?, and 3) What are the results of applying machine learning algorithms in the research on named-entity recognition?

2. Method

A literature study was conducted to identify and map the results of previous studies related to certain literature themes. In addition, a good literature study will produce a map of knowledge about a research topic that can guide researchers to dig deeper into areas that are not yet mature (Fisch & Block, 2018). The literature data in this study were collected through the use of Google Scholar with the keyword "Named Entity Recognition". The literature with the topic of introducing named-entity was then selected according to several factors, namely: 1) The approach used only focuses on the machine learning approach, and 2) The publication year of the literature obtained should be from the year 2018 to 2020. The results of the gradual selection resulted in seven pieces of literature which will be used as materials for comparison.

3. Results and Discussion

The process of analyzing and extracting large amounts of unstructured text or documents using Artificial Intelligence algorithms is often referred to as text mining. One part of text mining is the process of recognizing named-entities that can be used in various fields such as economy, health, social, politics, or culture. Based on the seven pieces of literature analyzed in this study, six pieces of literature apply the introduction of the main entity in the health sector, especially in the field of biomedicine and medicine. On the other hand, Wintaka's research used data taken from Twitter social media to identify the entity's name, location name, and organization name (Wintaka et al., 2019). The pieces of literature used in this research are shown in Table 1.

The health sector, especially the pharmaceutical industry, requires research on the introduction of named-entities, especially the medicine entities. The influence of a particular medicine with other medicines is closely monitored by the pharmaceutical industry in order to maintain patient safety from side effects caused by drug interactions (Chukwuocha et al., 2018). The biomedical field also has a very large corpus and requires information extraction to reduce the ambiguity due to several different entities that have the same acronym. Furthermore, several biomedical entities have inconsistent use of prefixes and suffixes (Cho et al., 2020).

Table 1. Literature Review Data based on Dataset

No	Ref	Year	Object / Dataset	Machine learning
1	(Chukwuocha et al., 2018)	2018	Medicine names / PubMed dataset	Conditional Random Field (CRF), and Naive Bayes (NB)
2	(Phan et al., 2019)	2019	Biomedical texts / BioNLP 2004 Challenge dataset	Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN)
3	(Casillas et al., 2019)	2019	Medical Online Corpus (GEN-MED) IXAMed Spanish EHR Corpus (EHR)	Bidirectional Long Short-Term Memory (Bi-LSTM), and Conditional Random Field (CRF)
4	(Suárez-Paniagua et al., 2019)	2019	eHealth-KD dataset	Bidirectional Long Short-Term Memory (Bi-LSTM), and Conditional Random Field (CRF)
5	(Wintaka et al., 2019)	2019	600 manually-labeled tweets in Bahasa Indonesia from Twitter social media	Bidirectional Long Short-Term Memory (Bi-LSTM), and Support Vector Machine (SVM)
6	(Gligic et al., 2019)	2019	Informatics for Integrating Biology & the Bedside – i2b2 dataset (2007-2012)	Forwards Neural Network (FFN), and Recurrent Neural Network (RNN), and Bidirectional Long Short-Term Memory

				(Bi-LSTM)
7	(Cho et al., 2020)	2020	JNLPBA (Biomedical) dataset NCBI (National Center for Biotechnology Information) dataset	Convolutional Neural Network (CNN), and Bidirectional Long Short-TermMemory (Bi-LSTM)

From Table 1, the popular machine learning algorithm used is the Conditional Random Field (CRF), while the popular deep learning algorithm is Bidirectional Long Short-Term Memory (Bi-LSTM). CRF is one of the various algorithms that are known to be great in building predictive models. CRF with its probabilistic model can be used for pattern recognition because it can consider word order labels that form sentences to identify entities from a text (Casillas et al., 2019). The LSTM algorithm is a development of the Recurrent Neural Network (RNN) algorithm through the generation of a memory cell that functions as a container for information for a long period. As for the Bi-LSTM algorithm, it has two layers that can move forward and backward. Bi-LSTM algorithm is generally used to handle sequential data to improve prediction accuracy (Cho et al., 2020).

The combination of Bi-LSTM and CRF approach is shown in Figure 1 (a), it shows two modules that compose a two-stage information extraction system. The input for the first Bi-LSTM layer is word embedding, in which the obtained output from the first layer is combined with word embeddings and sense-disambiguous embeddings in the second layer. Additionally, CRF was used in the final stage to get the most appropriate label for each token (Suárez-Paniagua et al., 2019). The concept of Medical Entity Recognition (MER) as shown in Figure 1 (b), relates to natural language processing applied to the clinical domain. The combination of Bi-LSTM with CRF serves to adapt the sequential tagger and to make it tolerant of high lexical variability and a limited number of corpus (Casillas et al., 2019).

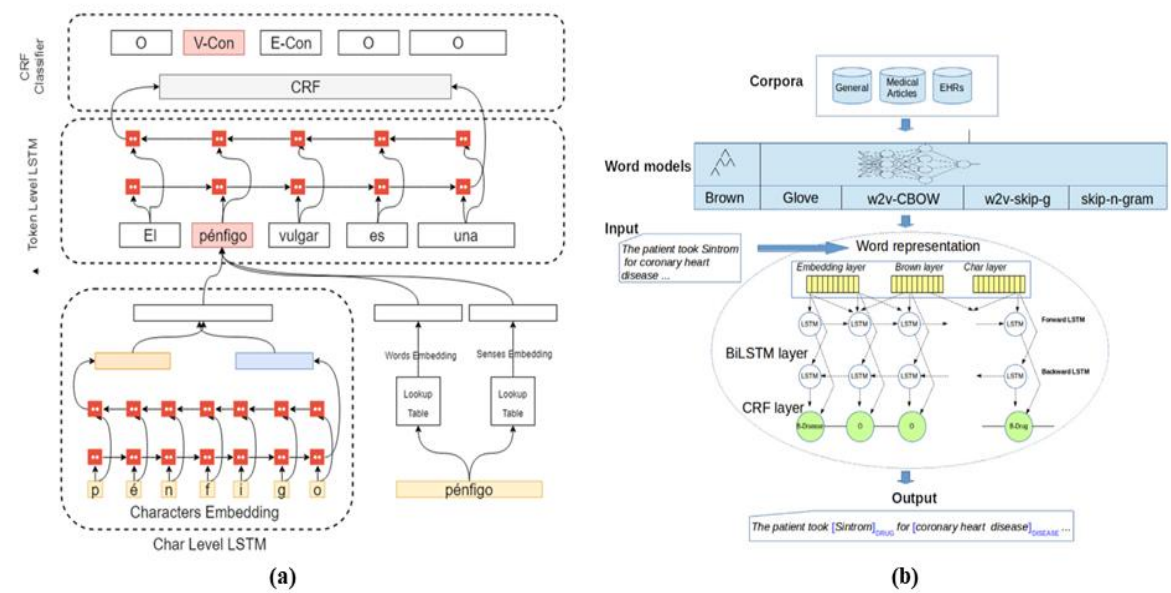


Fig. 1. (a) Two-stage deep learning approach (Casillas et al., 2019); (b) Neural architecture based on Bi-LSTM and CRF (Suárez-Paniagua et al., 2019);

A similar NER model but applied to different natural languages is still a frequent problem and it is still necessary to embed different trained words for each different natural language. As the first step, choosing the right algorithm and continuing to choose the corpus domain and genre is crucial to the success of the research.

Behind the various advantages of Bi-LSTM, there is still a problem where the complex Bi-LSTM algorithm architecture becomes one of the high computational burdens when applied to large-scale cases.

4. Conclusions

Literature studies of the previous researches obtained seven pieces of literature published from 2018 to 2020 with the research theme of the introduction of named-entity that can use 3 approaches, namely: 1) Machine learning, 2) Deep learning, and 3) A combination of machine learning with deep learning. The combination of machine learning and deep learning was used in 4 studies from 7 analyzed pieces of literature, namely the combination of Bidirectional Long Short-Term Memory (Bi-LSTM) deep learning with Conditional Random Field (CRF) machine learning. CRF with its probabilistic model can be used for pattern recognition because it is able to consider word order labels while the two layers of Bi-LSTM can handle sequential data to improve prediction accuracy by moving forward and backward. Currently, we are interested in exploring the topic of fake news detection but have not conducted any specific experiments related to NER. After conducting this literature study, we plan to explore NER using a fake news dataset that researchers have not been done before.

Acknowledgements

The authors wish to thank the Faculty of Engineering and Computer Science Universitas Komputer Indonesia for technical support. The Research presented in this paper has been done in the Laboratory of Accounting Information Systems, Universitas Komputer Indonesia.

References

Casillas, A., Ezeiza, N., Goenaga, I., Pérez, A., & Soto, X. (2019). Measuring the effect of different types of unsupervised word representations on Medical Named Entity Recognition. *International Journal of Medical Informatics*, *129*, 100–106. https://doi.org/10.1016/j.ijmedinf.2019.05.022

Cho, M., Ha, J., Park, C., & Park, S. (2020). Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*, *103*, 103381. https://doi.org/10.1016/j.jbi.2020.103381

Chukwuocha, C., Mathu, T., & Raimond, K. (2018). Design of an interactive biomedical text mining framework to recognize real-time drug entities using machine learning algorithms. *Procedia Computer Science*, *143*, 181–188. https://doi.org/10.1016/j.procs.2018.10.374

Fisch, C., & Block, J. (2018). Six tips for your (systematic) literature review in business and management research. *Management Review Quarterly*, *68*(2), 103–106. https://doi.org/10.1007/s11301-018-0142-x

Giarsyani, N. (2020). Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan. *Indonesian Journal of Applied Informatics*, *4*(2), 138. https://doi.org/10.20961/ijai.v4i2.41317

Gligic, L., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2019). Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *ArXiv*, *121*, 132–139.

Phan, R., Luu, T. M., Davey, R., & Chetty, G. (2019). Biomedical named entity recognition based on hybrid multistage cnn-rnn learner. *Proceedings - International Conference on Machine Learning and Data Engineering, ICMLDE 2018*, 136–141. https://doi.org/10.1109/iCMLDE.2018.00032

Suárez-Paniagua, V., Rivera Zavala, R. M., Segura-Bedmar, I., & Martínez, P. (2019). A two-stage deep

learning approach for extracting entities and relationships from medical texts. *Journal of Biomedical Informatics*, *99*, 103285. https://doi.org/10.1016/j.jbi.2019.103285
Wibisono, Y., & Khodra, M. L. (2018). Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin. *INA-Rxiv*. https://doi.org/10.31227/osf.io/vud2p
Wintaka, D. C., Bijaksana, M. A., & Asror, I. (2019). Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF. *Procedia Computer Science*, *157*, 221–228. https://doi.org/10.1016/j.procs.2019.08.161

Pendiskritan Model Lotka-Volterra menggunakan Kaedah Berangka Tidak Piawai dengan Trimean

Discretization of Lotka-Volterra Model using Nonstandard Finite Difference Scheme with Trimean

Noor Ashikin Othman[a]*, Mohammad Khatim Hasan[b], Bahari Idrus[c]

[a,b,c]Pusat Penyelidikan Teknologi Kecerdasan Buatan (CAIT), Fakulti Teknologi dan Sains Maklumat , Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
*Email: p97180@siswa.ukm.edu.my

Abstrak

Tujuan kajian ini untuk meramalkan nilai dua pemboleh ubah dalam ekonomi Malaysia iaitu keluaran dalam negara kasar (KDNK) dan harga minyak mentah (HMM). Pendiskritan model Lotka-Volterra berdasarkan kaedah berangka tak piawai dengan trimean (KBTPDT) dibangun untuk meramalkan nilai KDNK dan HMM. Perbandingan keputusan simulasi yang diperolehi dengan kaedah yang sedia ada iaitu kaedah berangka biasa (KBB) menunjukkan bahawa KBTPDT boleh meramalkan nilai dua pemboleh ubah dalam ekonomi Malaysia dengan ketepatan tinggi berdasarkan peratusan min ralat mutlak (MAPE).

Kata kunci: kaedah berangka tidak piawai; trimean; model Lotka-Volterra; keluaran dalam negara kasar; harga minyak mentah

Abstract

The aim of this study is to predict the values of two variables in the Malaysia economy which are gross domestic product (GDP) and crude oil price (COP). A non-standard approximation scheme with trimean approach is used to estimate the value of the GDP and COP based on Lotka-Volterra model. Comparisons of the obtained results with the standard method, demonstrate that non-standard approximation scheme with trimean method is able to predict the values of these variables effectively under the criterion of the minimization of mean absolute percentage error (MAPE).

Keywords: non-standard approximation scheme; trimean; Lotka-Volterra model; gross domestic product; crude oil price

1. Pendekatan Utama

Konsep KBTPDT telah diperkenalkan oleh (Hasan et al., 2018; Othman et al., 2019; Othman & Hasan, 2017) bagi mensimulasi model interaksi Lotka-Volterra, model Rosenzweig-Mac Arthur dan model Beddington-DeAngelis. KBTPDT telah dibangun untuk mengatasi masalah ketidakstabilan berangka yang wujud dalam kaedah berangka biasa (KBB). Antara KBB yang popular ialah kaedah Adam-Moultan dan kaedah Runge-Kutta(Hasan et al., 2015). Antara kelebihan utama KBTPDT ialah ia dapat meningkatkan ketepatan simulasi terhadap tingkah laku interaksi antara spesies dan dapat meramal titik keseimbangan dengan cepat dan tepat.

Keluaran dalam negara kasar (KDNK) dan harga minyak mentah (HMM) merupakan antara penentu ukur bagi perkembangan ekonomi (Mohd Saudi et al., 2019; Shrestha & Bhatta, 2018). Dalam kajian ini, konsep KBTPDT dibangun dalam model interaksi Lotka-Volterra untuk meramal nilai dua pemboleh ubah dalam ekonomi Malaysia iaitu KDNK dan HMM, kemudian dibandingkan dengan kaedah yang sedia ada iaitu KBB. KBB yang digunakan dalam kajian ini ialah kaedah Adam-Moultan. Perbandingan keputusan simulasi KBTPDT

yang diperoleh dengan KBB menunjukkan bahawa KBTPDT boleh meramalkan nilai pemboleh ubah ini dengan lebih jitu.

Model interaksi Lotka-Volterra adalah satu sistem persamaan pembezaan yang menggambarkan interaksi persaingan dinamik antara dua spesies. Model Lotka-Volterra dari dua spesies, x dan y, mengikut (Othman et al., 2020) adalah seperti berikut:

$$\frac{dx}{dt} = a_1 x(t) - b_1 x(t)^2 - c_1 x(t) y(t) \quad (1)$$

$$\frac{dy}{dt} = a_2 y(t) - b_2 y(t)^2 - c_2 y(t) x(t) \quad (2)$$

yang x dan y ialah saiz populasi spesies x dan y, a_i ialah kadar pertumbuhan spesies apabila ia hidup sendirian, b_i ialah had keupayaan niche untuk spesies dan c_i ialah hubungan interaksi antara dua spesies dengan $i = 1,2$.

Untuk membangun KBTPDT bagi persamaan (1), $\frac{dx}{dt}$ digantikan dengan $\frac{x_{i+1}-x_i}{\emptyset}$ mengikut (Mickens, 2002; Othman & Hasan, 2017). Dalam kajian ini fungsi pembahagi, $\emptyset$ adalah mengikut(Zibaei & Namjoo, 2014). Bagi sebelah kanan persamaan (1), nilai x, x^2 dan xy diganti dengan nilai anggaran masing-masing seperti:

$$x = 2x - x = 2x_i - x_{i+1}$$

$$x^2 = x_k x_{k+1} = x_i x_{i+1}$$

$$xy = x_{i+1} y_i$$

Maka persamaan (1) menjadi

$$\frac{x_{i+1} - x_i}{\emptyset} = A(2x_i - x_{i+1}) - B(x_i x_{i+1}) - C(x_{i+1} y_i)$$

$$x_{i+1} = \frac{A\emptyset 2x_i + x_i}{(1 + A\emptyset + B\emptyset x_i + C\emptyset y_i)} \quad (3)$$

Oleh kerana kaedah Trimean memerlukan dua nod sebelum untuk mengira nod seterusnya, maka persamaan (3) digunakan pada dua nod pertama. Persamaan trimean $\frac{x_{i-1}+2x_i+x_{i+1}}{4}$ diganti dalam nilai x, tanpa mengubah nilai x^2 dan xy, persamaan (1) menjadi

$$\frac{x_{i+1} - x_i}{\emptyset} = A\left(\frac{x_{i-1} + 2x_i + x_{i+1}}{4}\right) - B(x_i x_{i+1}) - C(x_{i+1} y_i)$$

$$x_{i+1} = \frac{A\emptyset 0.25 x_{i-1} + A\emptyset 0.5 x_i + x_i}{(1 - A\emptyset 0.25 + B\emptyset x_i + C\emptyset y_i)} \quad (4)$$

Untuk membangun KBTPDT bagi persamaan (2), $\frac{dy}{dt}$ akan digantikan dengan $\frac{y_{i+1}-y_i}{\emptyset}$ mengikut (Mickens, 2002; Othman & Hasan, 2017). Manakala bagi sebelah kanan persamaan (2), nilai y, y^2 dan xy digantikan dengan nilai anggaran masing-masing seperti

$$y = -y + 2y = -y_i + 2y_{i+1}$$

$$y^2 = y_k y_{k+1} = y_i y_{i+1}$$

$$xy = 2xy - xy = 2x_{i+1}y_i - x_i y_{i+1}$$

Maka persamaan (2) menjadi

$$\frac{y_{i+1} - y_i}{\emptyset} = P(-y_i + 2y_{i+1}) - Q(y_i y_{i+1}) - R(2x_{i+1}y_i - x_i y_{i+1})$$

$$y_{i+1} = -\frac{P\emptyset y_i - R\emptyset 2x_{i+1}y_i + y_i}{(1 - P\emptyset 2 + Q\emptyset y_i - R\emptyset x_i)} \quad (5)$$

Oleh kerana kaedah Trimean memerlukan dua nod sebelum untuk mengira nod seterusnya, maka persamaan (5) digunakan pada dua nod pertama. Persamaan trimean $\frac{y_{i-1}+2y_i+y_{i+1}}{4}$ digantikan dalam nilai y, tanpa mengubah nilai y^2 dan xy, persamaan (1) menjadi

$$\frac{y_{i+1} - y_i}{\emptyset} = P\left(\frac{y_{i-1} + 2y_i + y_{i+1}}{4}\right) - Q(y_i y_{i+1}) - R(2x_{i+1}y_i - x_i y_{i+1})$$

$$y_{i+1} = \frac{P\emptyset 0.25 y_{i-1} + P\emptyset 0.5 y_i - R\emptyset 2x_{i+1}y_i + y_i}{(1 - P\emptyset 0.25 + Q\emptyset y_i - R\emptyset x_i)} \quad (6)$$

Dalam kajian ini, model Lotka-Volterra yang didiskritkan dengan KBTPDT adalah persamaan (3) hingga persamaan (6) diguna untuk meramal KDNK dan HMM. Untuk memeriksa kejituan peramalan KBTPDT, peramalan KDNK dan HMM akan dibandingkan dengan data sebenar mengikut (Othman et al., 2020) . Kajian ini menggunakan peratusan min ralat mutlak (MAPE) (Othman et al., 2020) bagi menilai kejituan. Tafsiran nilai MAPE adalah kurang daripada 10% adalah ramalan yang sangat tepat manakala 10% hingga 20% adalah ramalan yang baik. Keputusan simulasi (dipaparkan dalam Jadual 1) menunjukkan bahawa KBB dan KBTPDT dapat meramalkan nilai KDNK dan HMM dengan jitu. Namun, KBTPDT dapat meramal KDNK dan HMM dengan lebih tepat berdasarkan nilai MAPE yang kurang daripada 10. Ini menunjukkan penyesuaian KBTPDT adalah lebih menghampiri kepada nilai sebenar berbanding dengan KBB.

Jadual 1. Simulasi nilai Keluaran Dalam Negara Kasar (KDNK) dan Harga Minyak Mentah (HMM) mengguna KBTPDT

	KDNK			**HMM**		
Tahun	Nilai Sebenar	KBB	KBTPDT	Nilai Sebenar	KBB	KBTPDT
2009	8559.23	8559.2634	8559.23	64.13	50.984761	64.13
2010	9040.57	9040.5988	9040.5791	79.64	118.4254	79.301019
2011	9372.01	9372.0434	9372.0347	99.91	105.13597	72.075941
2012	9743.10	9743.1105	9743.1115	101.58	84.404617	109.01285
2013	10061.72	10061.727	10061.751	99.19	82.548472	101.63637
2014	10524.07	10524.086	10524.07	91.40	88.578935	92.577627
2015	10912.15	10912.175	10912.153	53.51	52.932118	54.44564
2016	11219.63	11219.632	11219.644	46.84	55.961447	59.807457

2017	11720.74	11720.761	11720.768	55.91	65.017566	60.617053
2018	12109.49	12109.49	12109.49	69.78	69.531728	72.581455
MAPE		**0.0002**	**0.0002**		**14.8401**	**8.1223**

2. Kesimpulan

Kajian ini menggunakan pendiskritan KBTPDT untuk meramal nilai dua pemboleh ubah dalam ekonomi Malaysia bagi model interaksi Lotka-Volterra. Dua pemboleh ubah yang dipilih adalah KDNK dan HMM. Terdapat dua kaedah yang digunakan untuk mensimulasi interaksi antara KDNK dan HMM iaitu KBB dan KBTPDT. Hasil perbandingan keputusan simulasi yang diperolehi menunjukkan bahawa KBTPDT dapat meramalkan nilai dua pemboleh ubah ini dengan lebih tepat berbanding KBB. Ini menunjukkan pendiskritan KBTPDT lebih berkesan dan boleh dilaksanakan dalam mengkaji isu ekonomi. Kajian mengenai hubungan antara pemboleh ubah ekonomi adalah sangat penting dan berguna kepada kerajaan dalam membuat dasar-dasar pembangunan sosial untuk mensejahterakan ekonomi negara.

Penghargaan

Artikel ini adalah sebahagian daripada Geran Galakan Penyelidik [GGP-2017-023] Universiti Kebangsaan Malaysia.

Rujukan

Hasan, M. K., Abdul Karim, S. A., & Sulaiman, J. (2015). Graphical Analysis of Rosenzweig-MacArthur Model via Adams-Moultan and Fourth Order Runge-Kutta Methods. The 5th International Conference on Electrical Engineering and Informatics, 670–675.

Hasan, M. K., Othman, N. A., Karim, S. A. A., & Sulaiman, J. (2018). Semi non-standard trimean algorithm for Rosenzweig-MacArthur interaction model. International Journal on Advanced Science, Engineering and Information Technology, 8(4–2), 1520–1527. https://doi.org/10.18517/ijaseit.8.4-2.6779

Mickens, R. E. (2002). Nonstandard finite difference schemes for differential equations. Journal of Difference Equations and Applications, 8(9), 823–847. https://doi.org/10.1080/1023619021000000807

Mohd Saudi, N. S., Tsen, W. H., Murshidi, M. H., Harun, A. L., & Saayah, A. (2019). The Impact of Crude Oil, Natural Gas and Liquefied Natural Gas (LNG) Prices on Malaysia GDP: Empirical Evidence using ARDL bound Testing Approach. International Journal of Academic Research in Business and Social Sciences, 9(6), 988–1001. https://doi.org/10.6007/ijarbss/v9-i6/6060

Othman, N. A., & Hasan, M. K. (2017). New hybrid two-step method for simulating lotka-volterra model. Pertanika Journal of Science and Technology, 25(S6), 115–124.

Othman, N. A., Hasan, M. K., & Idrus, B. (2020). The crude oil price power on Malaysia GDP: Relationship Analysis employing Numerical Method with Optimisation Approach. Test Engineering and Management, 83(1101), 1101–1107.

Othman, N. A., Hasan, M. K., & Idrus, B. (2019). ALGORITMA BERANGKA TIDAK PIAWAI DENGAN TRIMEAN BAGI MODEL INTERAKSI BEDDINGTON-DEANGELIS. Kolokium Pendidikan (KOLOPEN), Institut Pendidikan Guru Kampus Perempuan Melayu, Melaka, 230–252.

Shrestha, M. B., & Bhatta, G. R. (2018). Selecting appropriate methodological framework for time series data analysis. The Journal of Finance and Data Science, 4(2), 71–89. https://doi.org/10.1016/j.jfds.2017.11.001

Zibaei, S., & Namjoo, M. (2014). A NSFD scheme for Lotka–Volterra food web model. Iranian Journal of Science and Technology, 38(4), 399–414. https://doi.org/10.22099/ijsts.2014.2556

Principal Component Analysis Variant Initialization in Convolutional Neural Network

Nor Sakinah Md Othman[a]*, Azizi Abdullah[b]

[a,b] Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
*Email: p108639@siswa.ukm.edu.my

Abstract

In Convolutional Neural Network (CNN), there are various weight initialization strategies that have been proposed to handle overfitting and slow convergence. This paper proposed an alternative weight initialization technique that utilizes Gaussian Principal Component Analysis (GPCA) initialization and Generalized Gaussian Principal Component Analysis (G-GPCA) initialization on LeNet-5 and AlexNet. The set of overlapping Gaussian windows is used to generate GPCA filters that simulates the characteristics of orientation and texture receptive fields which will lead to better performance in extracting low level features. The proposed method is tested on five different datasets namely MNIST, CIFAR-10, SVHN, GTSRB dan Covid-19. The results show that PCA variant initialization (PCA, GPCA, G-GPCA) obtained consistent accuracy that can be translated to consistent performance on a variety of datasets.

Keywords: Weight initialization; principal component analysis; convolutional neural network; image classification.

1. Introduction

Applications of convolutional neural network (CNN) on various domains such as object recognition have been increased significantly due to greater computing power and higher volume of training datasets. Even though it is a well-known research topic and various works have been introduced, the issue on obtaining consistent accuracy and faster convergence still persists. Several methods have been proposed to optimize the training process of CNNs and one of the approaches made by other researchers is by focusing on the weight initialization strategy (Koturwar & Merchant, 2017).

Popular methods of setting weights in the convolutional layer are Xavier initialization (Glorot & Bengio, 2010) and He initialization (He et al., 2015). In (Koturwar & Merchant, 2017), both methods are believed to be able to handle gradient diffusion problem and the dying neuron problem. Gradient diffusion can be defined as amplification or attenuation of gradient values throughout the backpropagation process which results to an event of exploding or vanishing of gradients (Sun, 2020). Utilizing the standard initialization has several downsides such as independent to data statistics (Koturwar & Merchant, 2017), prone to dying neuron problem (Lu et al., 2019) and are often produced in redundance (Luan et al., 2018). Thus, it motivates other different types of weight initialization techniques such as PCA initialization, Linear Discriminant Analysis (LDA) initialization (Alberti et al., 2017) and have shown promising results in image classification.

In this paper, initialization technique using Principal Component Analysis (PCA) variants such as PCA, Gaussian PCA (GPCA) and Generalized GPCA (G-GPCA) is introduced. LeNet-5 and AlexNet models are then used to investigate in CNN for image classification tasks. Several papers have introduced the usage of PCA filters on CNN, but there is still insufficient research conducted on GPCA and G-GPCA filters (Brause et al., 1999). In this method, PCA filters are generated by utilizing the image data statistics (Koturwar & Merchant, 2017). It has some advantages such as ability to handle gradient diffusion problem (Ren et al., 2016) and provide robustness against image transformations (Soon et al., 2020). The contribution of this work is to provide an

alternative in selecting a suitable weight initialization strategy that can produce consistent results over the wide range of datasets.

2. Related Works

The rise of research papers discussing weight initialization technique have led to various introduction of image filters in the deep learning environment. Many initialization techniques have been proposed such as zeros or identity initialization and random initialization (Li et al., 2020). Due to problems faced when applying the standard weight initialization strategy, this motivates other researchers to investigate other image filters such as PCA filter and LDA filter. LDA differs from PCA where, LDA aims to look for the components that maximize the class separation which requires class labels while, PCA aims to extract components that maximize the variance in the dataset (Alberti et al., 2017).

PCA initialization network has been thoroughly investigated by various works such as PCANet proposed by (Chan et al., 2015). Other works, such as 2-dimensional PCA weight initialization with parametric equalization normalization by (Wang et al., 2020), implementing PCA-initialized LeNet-5 model for image classification (Ren et al., 2016) and vehicular image classification with PCA CNN proposed by (Soon et al., 2020). With the increasing number of PCA initialization in deep learning, this shows promising future in using PCA filters in the convolutional layer. In this work, variant of PCA filters will be introduced and applied as an initialization strategy on LeNet-5 model and AlexNet model. The results showed that, PCA variant initialization is able to obtain consistent results and is highly robust towards any changes in the employed image dataset.

3. PCA Variant Initialization

The generated PCA variant filters will be inserted in the first convolutional layer of LeNet-5 and AlexNet model which will then be evaluated on numerous image datasets. Suppose that there are M input training images, $\{T_i\}_{i=1}^{M}$ of dimension $w \times h$ and the patch or sliding window size is $r \times r$.

3.1. Principal Component Analysis (PCA) filters

The process of generating PCA filter will utilize rectangular window, $F_{i,j}$ and it will be moved over the test image B $(w \times h)$ with a pre-defined step horizontally and then vertically to obtain sub-images, $U_{i,j}$. The test image B $(w \times h)$ at horizontal position $x \in \{0, \dots, w-1\}$ and vertical position $y \in \{0, \dots, h-1\}$ while, the sliding window $F_{i,j}$ $(r \times r)$ at position $i \in \{0, \dots, w-1\}$ and position $j \in \{0, \dots, h-1\}$.

The weighting of the pixels for each sub-image, $U_{i,j}$ is a product of multiplying the pixel values and the specific window weights, $F_{i,j}$ at the associated window locations. The window weights of rectangular window $F_{i,j}$ and corresponding sub-image $U_{i,j}$ can be obtained based on equation (1) and equation (2) respectively,

$$F_{i,j}(x,y) = \begin{cases} 1 \; if \; 0 \le x - i < r \; and \; 0 \le y - j < r \\ 0 \; else \end{cases} \tag{1}$$

$$U_{i,j}(x,y) = F_{i,j}(x,y) \cdot B(x,y) \tag{2}$$

The collected non-overlapping sub-images of the i-th image, $x_{i,1}, x_{i,2}, \dots, x_{i,\hat{w}\hat{h}} \in \mathbb{R}^{r \times r}$ where each $x_{i,j}$ denotes the j-th vectorized patch in T_i and $\hat{w} = w - \left(\frac{r}{2}\right), \hat{h} = h - \left(\frac{r}{2}\right)$. Each image vector is then normalized by subtracting its respective mean to ensure that the image vector is centered which results to, $\bar{X} = \left(\bar{x}_{i,1}, \bar{x}_{i,2}, \dots, \bar{x}_{i,\hat{w}\hat{h}}\right)$ where each $\bar{x}_{i,j}$ is a mean removed patch. By constructing the same matrix for all input images and assemble them together will results to,

$$\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_M) \in \mathbb{R}^{rr \times M\hat{w}\hat{h}} \tag{3}$$

The PCA filters are then expressed as follows, assuming that there are n principal eigenvectors,

$$W_l^1 = mat_{r \times r}\left(q_l(\bar{X}\bar{X}^T)\right) \in \mathbb{R}^{r \times r}, l = 1,2, \dots, n \tag{4}$$

Where $mat_{r\times r}(v)$ is a function that maps $v \in \mathbb{R}^{r\times r}$ and $q_l(\bar{X}\bar{X}^T)$ denotes the l-th principal eigenvector of $\bar{X}\bar{X}^T$.

Production of PCA filters initially can only accommodate images with single input (i.e., grayscale image) thus, multi-channel PCA filters is produced to cater for colored input images (Chan et al., 2015). This is essential if the filters generated are required to capture complex features from complex image database. Similar to the existing step of constructing image matrix X, an additional individual matrix of each RGB channel is created and is denoted by $X_R, X_G, X_B \in \mathbb{R}^{rr\times M\hat{w}\hat{h}}$, respectively. Production of PCA filters for RGB channels are then similar to equation (4) with a slight variation which is defined as follows,

$$W_l^{R,G,B} = mat_{r\times r\times 3}\left(q_l\left(\tilde{X}\tilde{X}^T\right)\right) \in \mathbb{R}^{r\times r\times 3}, l = 1,2,\dots,n \tag{5}$$

where $\tilde{X} = \left[{X_R}^T, {X_G}^T, {X_B}^T\right]^T$ and $mat_{r\times r\times 3}(v)$ is a function that maps $v \in \mathbb{R}^{r\times r\times 3}$ into a matrix $W \in \mathbb{R}^{r\times r\times 3}$.

3.2. Gaussian PCA (GPCA) filters

Generation of GPCA filters are similar to the generation of PCA filters with a slight variation to the sliding window used which is the Gaussian window, $G_{i,j,\sigma}$. The Gaussian window weights can be defined as equation (6) below,

$$G_{i,j,\sigma}(x,y) = F_{i,j}(x,y)\ \cdot \exp\left(-\frac{1}{2\sigma^2}\cdot\left[\left(x-i-\frac{r-1}{2}\right)^2 + (y-j-\frac{r-1}{2})^2\right]\right) \tag{6}$$

In comparison to the rectangular window, Gaussian window are weighted with a two-dimensional Gaussian function and is not weighted in unity. Thus, the gaussian parameter, σ represents the width of the Gaussian function. The Gaussian window weights, $G_{i,j,\sigma}$ will be used to obtain the overlapping sub-image $U_{i,j}$ by multiplying the pixel value of test image B. The remaining process in generating GPCA filters are the same as the process of producing PCA filters from equation (3) until equation (5).

3.3. Generalized GPCA (G-GPCA) filters

Generation of PCA and GPCA filters are required to be recalculated for each image which is a strenuous process. However, this can be handled by approximating and defining a generalized equation for GPCA. This enhancement provides an advantage of creating filters that are suitable for different images by only setting the correlation parameters and gaussian parameter. The generalized GPCA (G-GPCA) filters can be defined as follows,

$$\sum_{z'=0}^{r-1} \sigma_B \cdot \exp\left[-\frac{1}{2}\cdot\sigma^{-2}\cdot\left(\left(z-\frac{r-1}{2}\right)^2+\left(z'-\frac{r-1}{2}\right)^2\right)\right]\cdot\exp(-\gamma\cdot|z-z'|)\cdot q_\gamma(z') = \lambda_\gamma\cdot q_\gamma(z) \tag{7}$$

Based on equation (7), GPCA filters is a product of two one dimensional eigenvectors $q_\alpha(x)$ and $q_\beta(y)$. Analytical formulation of the eigenvectors can be formed as follows,

$$q_\gamma(z) = a\cdot\cos\left(b\cdot\gamma\cdot\left(z-\frac{r-1}{2}\right)\right) \text{ and } \qquad q_\gamma(z') = a'\cdot\sin\left(b'\cdot\gamma\cdot\left(z'-\frac{r-1}{2}\right)\right) \tag{8}$$

Where, the parameter a and a' represents constant parameter while, parameters b and b' can be calculated as follows,

$$b\cdot\tan\left(\frac{1}{2}\cdot b\cdot\gamma\cdot r\right) = 1 \quad \text{and} \qquad b'\cdot\cot\left(\frac{1}{2}\cdot b'\cdot\gamma\cdot r\right) = -1 \tag{9}$$

Moreover, calculating the eigenvalues can be conducted as follows,

$$\sigma[B(x,y)]\cdot\sigma[B(x',y')] = \sigma_B\cdot g_{r,\sigma}(x,y)\cdot g_{r,\sigma}(x',y') \tag{10}$$

$$\text{where,} \qquad g_{r,\sigma}(x,y) = \exp\left[-\frac{1}{2}\cdot\sigma^{-2}\cdot\left(\left(x-\frac{r-1}{2}\right)^2+\left(y-\frac{r-1}{2}\right)^2\right)\right]$$

GPCA filters slightly differs due to the usage of Gaussian window, thus the product of the standard deviations is equals to,

$$\lambda_\gamma = \frac{2}{\gamma \cdot (b^2+1)} \quad \text{and} \quad \lambda_\gamma = \frac{2}{\gamma \cdot (b'^2+1)} \tag{11}$$

4. Experiments

4.1. Descriptions of Dataset

Datasets used for the LeNet-5 model are MNIST, CIFAR-10, SVHN and GTSRB dataset while, AlexNet model is the Covid-19 dataset (Cohen et al., 2020). The Covid-19 dataset contains 4 class category of Covid-19, normal, pneumonia bacterial and pneumonia virus which amounts to a total of 306 images (270 for training, 36 for testing). Meanwhile, MNIST dataset contains 10 class category and total of 70k images (60k for training, 10k for testing), CIFAR-10 dataset contains 60k images with 10 classes (50k for training, 10k for testing), SVHN dataset which comprises of 10-digit classes (73,257 for training, 26,032 for testing) and GTSRB dataset with 43 classes for traffic signs (39,209 for training, 12,630 for testing).

4.2. Experimental Setup

In this paper, the deep learning library for Java called IntelliJ was used to implement LeNet-5 and AlexNet model with variation of weight initialization strategy. The learning rate set for both models are 0.01 for CIFAR-10 dataset and 0.001 for the remaining datasets. Other hyperparameter values such as weight decay is set to 5×10^{-4}, momentum of 0.9 and epoch value is 100 for LeNet-5 model and 200 for the AlexNet model. The gaussian parameter set in the GPCA and G-GPCA filter generation process for LeNet-5 model is 0.8 while for the AlexNet model is 2.8. Both models are implemented on an Intel Core i7-10875H @ 2.3 - 5.1 GHz with Nvidia RTX2070 GPU of 16GB RAM.

4.3. Preliminary Result

PCA initialization model obtains the highest accuracy for MNIST, CIFAR-10 and GTSRB dataset while, GPCA performs the best for SVHN dataset. Based on Table 1, GPCA initialization achieves better accuracy than G-GPCA in CIFAR-10 and SVHN dataset but only with a slight margin of ($\approx$ 0.45%) and it achieves lower results than G-GPCA in MNIST, GTSRB and Covid-19 dataset with a difference of ($\approx$ 5.91%). The slight variation in accuracy shows that applying G-GPCA can gives similar accuracy to GPCA with faster filter generation process. PCA variant initialization shows promising results due to the consistency of achieving similar accuracy for each weight initialization strategy and datasets used. Further comparison with other weight initialization technique such as Xavier and Gabor will be conducted.

Table 1. Comparison of accuracy between PCA variant initialization on LeNet-5 and AlexNet model (Bolded value is the highest accuracy).

	LeNet-5				AlexNet
Filters	Accuracy (%)				
	MNIST	CIFAR-10	SVHN	GTSRB	Covid-19
PCA	**99.13**	63.97	88.81	92.77	63.89
GPCA	99.10	63.95	**89.24**	89.68	52.78
G-GPCA	99.12	63.23	89.07	90.72	**69.44**

Further work contains detailed and thorough experiments in analyzing the performance of PCA variant initialized network. Moreover, future work also comprises of experimenting and performing hybridization of PCA variant filters with Gabor filters.

Acknowledgements

We would like to thank the Ministry of Education who awarded us a FRGS/1/2019/ICT02/UKM/02/8 research grant entitled "Ensemble of Convolutional Neural Networks Using Multiple Heterogeneous Filter Models for Image Classification".

References

Alberti, M., Seuret, M., Pondenkandath, V., Ingold, R., & Liwicki, M. (2017). Historical document image segmentation with LDA-initialized deep neural networks. In Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (pp. 95-100).

Brause, R., Arlt, B., & Tratar, E. (1999). Project semacode: A scale-invariant object recognition system for content-based queries in images databases.

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification?. IEEE transactions on image processing, 24(12), 5017-5032.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:2006.11988.

Koturwar, S., & Merchant, S. (2017). Weight initialization of deep neural networks (dnns) using data statistics. arXiv preprint arXiv:1710.10570.

Li, H., Krček, M., & Perin, G. (2020). A Comparison of Weight Initializers in Deep Learning-Based Side-Channel Analysis. In International Conference on Applied Cryptography and Network Security (pp. 126-143). Springer, Cham.

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying relu and initialization: Theory and numerical examples. arXiv preprint arXiv:1903.06733.

Luan, S., Chen, C., Zhang, B., Han, J., & Liu, J. (2018). Gabor convolutional networks. IEEE Transactions on Image Processing, 27(9), 4357-4366.

Ren, X. D., Guo, H. N., He, G. C., Xu, X., Di, C., & Li, S. H. (2016). Convolutional neural network based on principal component analysis initialization for image classification. In 2016 IEEE first international conference on data science in cyberspace (DSC) (pp. 329-334). IEEE.

Soon, F. C., Khaw, H. Y., Chuah, J. H., & Kanesan, J. (2020). Semisupervised PCA Convolutional Network for Vehicle Type Classification. IEEE Transactions on Vehicular Technology, 69(8), 8267-8277.

Sun, R. Y. (2020). Optimization for deep learning: An overview. Journal of the Operations Research Society of China, 8(2), 249-294.

Wang, Y., Rong, Y., Pan, H., Liu, K., Hu, Y., Wu, F., ... & Chen, J. (2020). PCA Based Kernel Initialization for Convolutional Neural Networks. In International Conference on Data Mining and Big Data (pp. 71-82). Springer, Singapore.

Preliminary Work on Bag-of-Requirement Representation for SMA Reviews

Mustafa Abdulkareem[a*], Sabrina Tiun[b], Umi Asma` Mokhtar[c] , Masnizah Mohd[d]

[a,b]Asian Language Processing (ASLAN) Lab, CAIT, Faculty of Information Science and Technology, UKM, Selangor, Malaysia
[c]Information Governance (IG) Lab, CYBER,Faculty of Information Science and Technology, UKM, Selangor, Malaysia
[d]Cyber Intelligence (CI) Lab, CYBER, Faculty of Information Science and Technology, UKM, Selangor, Malaysia
*Email: P93823@siswa.ukm.edu.my

Abstract

Information extraction from user reviews has recently been used for social requirement retrieval to enhance the social media apps (SMA) development process. However, with the overwhelming amount of reviews, one needs to identify reviews that are relevant and ignore reviews that are not relevant or non-informative. To facilitate in identifying the relevant reviews, a representation related to social requirement terms of such review seems necessary. This paper describes our on-going work on building the SMA reviews representation by combining the Wordnet lexical and review word embedding and relates the representation to the standard social requirements. The work is still at the very early stage where we layout the method to combine the word embedding and Wordnet, expanding the social requirement term, and relating review to social requirements.

Keywords: Social requirement, Social media application, social media reviews, Word embedding, WordNet;

1. Introduction

Nowadays, social media applications (SMA) such as Facebook, Instagram, Twitter, etc., play a large part in people's lives, such as sharing feelings and their daily activities Continuously, which makes this information easy to access and in a simple way for anyone (Hemmatirad 2020). The sharing of information by the massive number of users has led to the emergence of many text classification techniques, opinion mining, and information extraction (Martinez-Rodriguez, Hogan, and Lopez-Arevalo 2020).

Popular application platforms such as Apple and Android allow users to evaluate applications by writing texts that refer to personal reviews about the application. These reviews are among the rich sources for software developers to upgrade their applications to users' needs. Also, these reviews are of high importance for new users.

2. Research Gap and Related Work

Existing research on mining app store reviews has been focused on extracting and classifying technically informative reviews into bug reports and feature requests (Alomar et al. 2021). However, little attention has been paid to extracting and synthesizing the social requirements present in user reviews. Social requirements, such as security, privacy, and performance, enforce various design constraints throughout the software development process (Whitworth 2011). Addressing these constraints is a critical factor for achieving user satisfaction and maintaining market practicality.

The following summarizes previous literature's limitations and provides several exciting research directions. There is an overwhelming amount of reviews; most of these reviews are not relevant or non-informative to the evaluating models. For example, SMA can have hundreds of thousands or even millions of reviews. Facebook SMA gets more than 10000 reviews every three days (Xiao et al. 2020). Challenges stem from the data's scale, unique format, diverse nature, and a high percentage of irrelevant information and spam (Häring, Stanik, and Maalej 2021).

In previous works, information extraction, keyword extraction, feature extraction, etc. method employed a vast, well-organized public lexicon known as WordNet to avoid a vast annotated corpus (Orkphol & Yang 2019). Recently, WordNet has been eclipsed by the success of the new lexical similarity benchmarks with the achievement of word embedding (Jimenez et al. 2019). WordNet's improvement by combining other word embedding Word2vec and Word2set has achieved better results than the classical WordNet-based approaches and competitive with those neural embeddings. The word relatedness affected by that combination makes the efficiency for development (Lee et al. 2019). Furthermore, extracting related reviews to social requirement terms using classical WordNet-based will not produce better results due to the weakness of word relatedness as the direct semantic relations that assuming the links between concepts represent distances. In addition, such links do not cover all possible relations between synsets. In this study, we present how to tackle the weakness of Wordnet in representing review by combining word embedding and Wordnet lexical that later can be used to extract SMA reviews that related to the only social requirement.

3. Bag-of-requirement (BOR) Representation

To handle the weakness of classic WordNet representation, this work proposes a word representation based on extended WordNet and word embedding named bag of requirement (BOR) representation. The BOR addresses the data sparseness and cuts off the threshold of classic WordNet representation. The method to build BOR of SMA reviews consists of three main steps, given the user reviews and social requirement term (SRT) as the input as shown in Figure 1.

- Step 1: Combine word embedding and Wordnet to expand WordNet.
- Step 2 Build SRTV by enriching SRT.
- Step 3: Build BOR using vectorized reviews (bag-of-word, unigram) and SRTV.

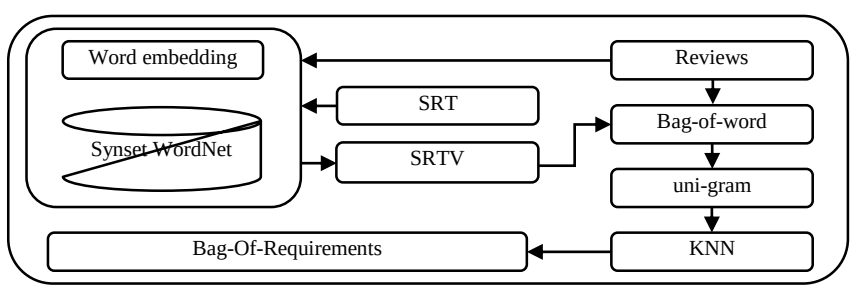


Fig. 1. Building BOR for SMA reviews

3.1. Step 1: Combining word embedding and wordnet

This step combines word embeddings and WordNet lexical database. Let Si, m be the m-the sense associated with the word wi. Then the path distance between the senses of all the noun pairs and some verb pairs can be

computed in WordNet. The final semantic relatedness score l of two words that combines WordNet and word embedding is defined as in Eq.1:

$$\mathrm{rel}(w_i, w_j) = \max_{m,n\,\lambda} \cos\left(v(w_i), v(w_j)\right) + (1-\lambda)\frac{1}{\mathrm{dist}(Sw_i, Sw_j)} \quad (1)$$

Where dist (Si, Sj) is the distance between two senses Si, m, and Sj, n. $v(w_i), v(w_j)$ are the vector representations of word wi and wj in the word embedding. λ is a weighting factor. The final combined word to combined sense representation model is defined as in Eq. 2:

$$W2CS(w) = \{w_j | w_j : \ \mathrm{rel}(w_i, w_j) > t_1\} \quad (2)$$

3.2. Step 2: SRT enrichment to build Social Requirement Terms Vocabulary SRTV

In this work, SynSet WordNet is used to enrich SRT by adding the synonymous words onto the existing keywords to achieve the Social Requirement Terms Vocabulary (SRTV). Eventually, it will increase the distinct gap between related reviews and non-related reviews in classifying the reviews of the applications. SRTV contains a group of words labeled to each term of SRT, as shown in Figure 2.

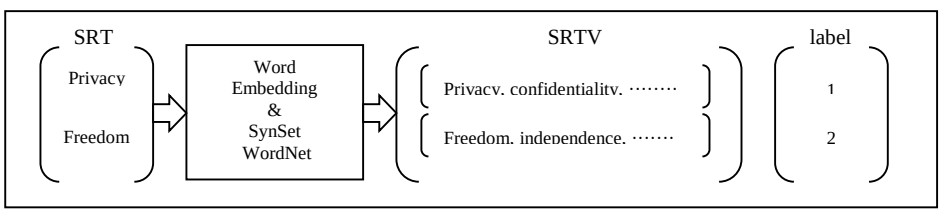


Fig. 2. Enriching SRT to build SRTV.

3.3. Step 3: Build BOR by relating SRTV label with SMA review

The Bag-of-Requirements or BOR (as shown in Figure 3) is a SMA review that has been vectorized into BOW and n-gram and labeled with SRTV. The semi-supervised KNN approach applied to the vectorized SRTV and reviews to achieve multi BOR labeling for each review.

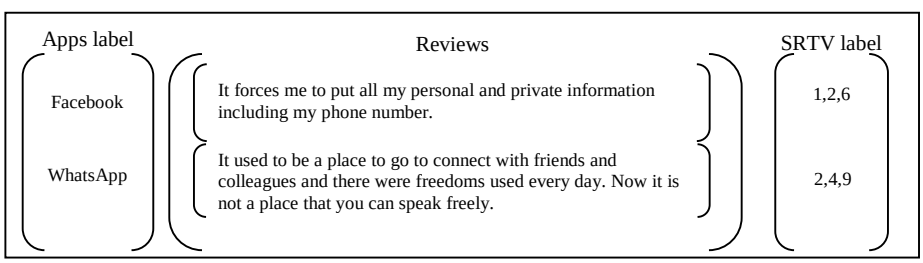


Fig. 3. Bag-of-Requirement data organization.

4. Conclusion

We proposed BOR in this paper, an approach that extracts social requirements of SMA reviews and then seeks matching requirements for each review. The combination of WordNet and word embedding has been approved to achieve preferable results due to the enhancement of the word relatedness in classical WordNet-based approaches. As we know, SMA developers receive huge numbers of reviews every day, making manual analysis difficult and time-consuming. Thus, the application of BOR that filters irrelevant reviews helps in cutting cost and speed up upgrading or improving the SMA. In addition, our proposed BOR also can be expanded to different applications, as such sentiment analysis, system recommendation, application evaluation, and so on.

Acknowledgments

This work is partially supported by MoHE under research code: FRGS/1/2020/ICT02/UKM/02/1

References

Alomar, Eman Abdullah, Wajdi Aljedaani, Mohamed Wiem Mkaouer, Yasmine Elglaly, Murtaza Tamjeed, William Catzin, Yasmine N. El-Glaly, and Yasmine N. El. 2021. "Finding the Needle in a Haystack: On the Automatic Identification of Accessibility User Reviews." 16. doi: 10.1145/3411764.3445281.

Häring, Marlo, Christoph Stanik, and Walid Maalej. 2021. "Automatically Matching Bug Reports With Related App Reviews."

Hemmatirad, Kimia. 2020. "Detection of Mental Illness Risk on Social Media through Multi-Level SVMs." 116–20.

Jimenez, Sergio, Fabio A. Gonzalez, Alexander Gelbukh, and George Duenas. 2019. "Word2set: WordNet-Based Word Representation Rivaling Neural Word Embedding for Lexical Similarity and Sentiment Analysis." *IEEE Computational Intelligence Magazine* 14(2):41–53. doi: 10.1109/MCI.2019.2901085.

Lee, Yang-Yin, Hao Ke, Ting-Yu Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. "Combining and Learning Word Embedding with WordNet for Semantic Relatedness and Similarity Measurement." *Journal of the Association for Information Science and Technology* 00(0):1–14. doi: 10.1002/asi.24289.

Martinez-Rodriguez, Jose L., Aidan Hogan, and Ivan Lopez-Arevalo. 2020. *Information Extraction Meets the Semantic Web: A Survey*. Vol. 11.

Orkphol, Korawit, and Wu Yang. 2019. "Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet." doi: 10.3390/fi11050114.

Price. 2019. "The Top 20 Social Media Apps and Sites in 2019." *Https://Www.Makeuseof.Com/Tag/Top-Social-Media-Apps-Sites/*.

Sharma, Chhavi, Minni Jain, and Ayush Aggarwal. 2018. "Keyword Extraction Using Graph Centrality and WordNet." Pp. 363–72 in *Towards Extensible and Adaptable Methods in Computing*. Springer Singapore.

Wei, T., Y. Lu, H. Chang, Q. Zhou, X. Bao-Expert Systems with Applications, and undefined 2015. n.d. "A Semantic Approach for Text Clustering Using WordNet and Lexical Chains." *Elsevier*.

Whitworth, Brian. 2011. "The Social Requirements of Technical Systems." *Virtual Communities* 1461–81. doi: 10.4018/978-1-60960-100-3.ch424.

Xiao, Jianmao, Shizhan Chen, Qiang He, Hongyue Wu, Zhiyong Feng, and Xiao Xue. 2020. "Detecting User Significant Intention via Sentiment-Preference Correlation Analysis for Continuous App Improvement." Pp. 386–400 in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12571 LNCS. Springer Science and Business Media Deutschland GmbH.

Towards on Comparing Conventional Query Expansion Approaches and Word Embedding-Based Approaches

Yasir Hadi Farhan[a]*, Shahrul Azman Mohd Noah[b], Masnizah Mohd[c]

[a,b,c] Faculty of Information Science and Technology, Uninvirsiti Kebangsaan Malaysia, Bangi 43000, Malaysia
*Email: yasir.hadi87@yahoo.com

Abstract

Automatic Query Expansion (AQE) is a popular way that is used to address the problem of term mismatch between the user's query terms and the relevant documents in the corpus. Term mismatch occurs when the user presents his/her query but it does not match the contents present in the existing documents. This can severely affect the search-based tasks. Researches have been proposed several approaches to solve this issue, such as involving new related terms or using synonyms of the given query. The proposed approaches can be divided into two main groups: Conventional Query Expansion Approaches, and Word Embedding-Based Approaches. Conventional approaches such as linguistic and ontology-based approaches have been proposed to address the vocabulary mismatch problem. Word Embedding (WE) has been recently used to address the problems mentioned above as exhibited by the term mismatch. Word2Vec is a WE toolkit that transfers the words existing in the vocabulary to vectors of the actual numbers. In this paper, we have reviewed and summarized the Conventional approaches and the approaches based on the Word Embedding for AQE.

Keywords: Automatic Query Expansion, Information Retrieval, Word Embedding;

1. Introduction

Web Searching is considered as one of the most prominent and valuable services on the Internet. It has many web documents that attract the users to get the useful information they are looking for through the search engine. The main purpose of an Information Retrieval (IR) structure is to retrieve documents that are most relevant to the user's query, and the most relevant documents are the best IR systems than those which are less relevant. The documents are ranked in terms of the query and terms of the retrieved documents (El,2020). The user must usually formulate the information requirements via a query; then the IR system returns the information to the user (Baeza-Yates & Ribeiro-Neto,1999). During interactions with users, IR systems face many challenges, one of which is vocabulary, also called vocabulary mismatch (Carpineto & Romano,2012; Farhan Yasir et al.,2020).

Several efforts have been made within the research community to improve the effectiveness of IR systems, including the use of relevance feedback and query refinement. Researchers in the field of IR have suggested many solutions to address this issue, the latest being the AQE. This technique is aimed at rephrasing the original query by adding new terms to improve the accuracy of the IR system (Abbache et al.,2016). Some of the proposed approaches rely on the feedback given by the users to expand their queries by inserting new related terms into the original query or suggesting appropriate keywords (Raza et al.,2018).

2. Automatic Query Expansion

Automatic Query Expansion (AQE) is reformulating the user's query by automatically adding additional relevant terms to the original query to improve retrieval performance. The fundamental issue of the retrieval process is the vocabulary mismatch between the query terms and the documents. Recently, several AQE approaches, such as linguistic and ontology-based approaches, have been proposed to address the vocabulary mismatch problem (Raza et al.,2019).

Nowadays, AQE methods are recommended as an efficient method in addressing the query-document words and terminology discrepancy problem in IR tasks (Vechtomova,2009; White & Horvitz,2015). The goal of AQE is to enhance retrieval performance by adding some semantical words to the original query. AQE approaches can be categorized as global or local approaches. Global approaches extend the original query independently regardless of the outcome. WordNet is typically the standard exogenous tool to choose from to select new terms semantically related to the original ones. (Pal et al.,2014). Local methods, by contrast, utilize approaches to Relevance Feedback (RF). The results of a first retrieved documents are used to select the most promising terms for the initial query. AQE can be categorized into two main groups: Conventional Query Expansion Approaches, and Word Embedding-Based Approaches, each group are explained in details in the next sections.

2.1. Conventional Query Expansion Approaches

As stated by Cui et al. (2002), to assist the web users in developing better queries or requests, researchers have concentrated on AQE approaches. In AQE, users provide extra input on query phrases or words by proposing additional query terms or words. Web search engines, such as Google and Yahoo, give a query suggestion to the users. Query suggestions are a common search experience that displays an updating list of relevant queries that users can select from as they type. These suggestions help users find specific queries guaranteed to have better results (Wang et al.,2009).

One of the conventional AQE techniques is to find the related words of the given query by using the thesaurus to pick the synonyms for that word. WordNet considered as the most popular methods (Jiang & Conrath,1996; Mandala et al.,1998). WordNet is a lexical database that groups words into a set of synonyms called synsets. This technique expands the original query by analyzing the expansion features such as lexical, morphological, semantic, and syntactic term relationships. Several studies were found to expand the query using the WordNet (Mahgoub et al.,2014; Al-Chalabi et al.,2015; Abbache et al.,2016).

Another AQE technique is Relevance Feedback (RF), where it is one of the most effective technique used to expand the users query, where the terms are extracted from the top retrieved documents. Pseudo Relevance Feedback (PRF) is the most similar technique to RF (ALMasri et al.,2016; Singh & Sharan,2017). PRF technique was proposed initially by (Croft & Harper,1979), where it assumes that the top retrieved documents are relevant, then select from these documents related terms to add to the original query. Some studies are using the PRF to expand the users queries (Atwan et al.,2016; El Mahdaouy et al.,2019).

One of the earliest AQE techniques is stemming. Stemming is the process of reducing the the inflected words to their morphological root or word stem. It combines the words have one stem (assuming they have the same meaning) to make them inder one index term. The stemming technique can be simple by removing pluralization suffixes from words or complicated ways of preserving meanings and incorporating dictionaries (Farrar & Hayes,2019). Few approaches that use stemming for AQE (Hammo et al.,2007; Khafajeh et al.,2010; Nwesri & Alyagoubi,2015). Co-occurrence of the words is considered as one of the main ways that compute the semantic relations between the words. The hypothesis is that, semantically similar words almost occur in the same contexts (Z,1968; Lindén & Piitulainen,2004). Shaalan et al. (2012) was used the co-occurrence of words for AQE.

2.2. Word Embedding-Based Approaches

Since the distributional hypothesis was proposed by (Harris,1954), large unlabeled text corpora have been often used to build word representations. Low dimensional representations know as Word Embeddings (WE) have recently resulted in low-dimensional representations, as the loss function usually using the algorithm of Stochastic Gradient Descent, often in the form of a neural network, is minimized (Mikolov et al.,2013; Pennington et al.,2014). These so-called WE have yielded state-of-the-art results in various NLP tasks such as word similarity, analogy, PoS tagging, named-entity disambiguation, or IR tasks (Collobert et al.,2011; Socher

et al.,2011; Mikolov et al.,2013). One drawback of the previous methods is that they operate at word-level, so that morphological rich words or vocabulary words can be modelled more closely. WE has been recently used to address the vocabulary mismatch problem (Roy et al.,2016; El Mahdaouy et al.,2018; Fernández-Reyes et al.,2018). WE are distributed representations techniques of the words commonly extracted from a neural network that models the joint distribution of the corpus vocabulary. The embedding models are usually trained in a broad corpus based on term proximity (Diaz et al.,2016).

Recently, most research in AQE relies on WE as a semantic modelling technique (ALMasri et al.,2016; Roy et al.,2016). To leverage WE to improve AQE effectiveness, Roy et al. (2016) proposed three AQE methods based on the WE technique. The AQE technique is devised using semantic relationships in a distribution of the terms, where the candidate related terms have been obtained using the K-Nearest Neighbor (K-NN) approach. Several studies were found to use the WE for AQE (Zamani & Croft,2016; Zamani & Croft,2017; El Mahdaouy et al.,2018).

El Mahdaouy et al. (2019) proposed incorporating WE similarity into PRF models for AIR. The principal idea is to select expansion terms in the PRF documents with their distribution and their similarity to the original query terms. The study hypothesizes that WE can be used for AIR in the PRF framework, as similar words to be grouped together to one side are close to each other in the vector space. The main aim is to increase the weight and the original query terms in semantically related terms. Three neural WE models, including a Skip-gram, Words CBOW Continuous Bag and Glove, are investigated in this work. Evaluations are carried out using three neural WE models on the standard TREC 2001/2002 Arabic test collection. The aim of the study is to understand how WE can be used in PRF techniques for AIR. Results showed that the PRF extensions proposed exceed the PRF baseline models significantly. In addition, they increased by 22% the basic IR model for MAP and 68% the robustness index. In addition, there was not statistically significant performance difference between the three model WE models (Glove, CBOW and Skip-gram).

Besides, Maryamah et al. (2019) proposed an AQE method based on *BabelNet* using the WE technique Word2Vec. *BabelNet* is a semantic search dictionary that combines knowledge of Wikipedia articles and lexicographic from Wordnet (Navigli & Ponzetto,2012). WordNet is used to acquire the synsets based on lexical or semantic relationships between terms, whereas uses the relationship between entities on the Wikipedia page. The candidate expansion term is also obtained from WordNet and synonyms. Based on the experiment results of 40 queries, the average accuracy is the study was 90%.

In addition, Wang et al. (2019) proposed a novel AQE method by using the K-NN method, where they utilize the local WE and focusing on the semantic similarity between the words. The cosine similarity measure is utilized to calculate the similarity between two words. Based on the experimental results, they demonstrate that the proposed local embedding method in significant outperforming the baselines methods and its promising area for future work in AQE. Finally, a comparison between the conventional AQE approaches and the WE-based approaches is given in the next paragraph. First, the input and the output of the conventional AQE approaches are terms, while the input of the WE-based approaches is terms and the output is vectors. Second, the complexity in the conventional AQE approaches depends on the nature of the language used, whereas the WE-based approaches deal with any Language easily because it relying on the vectors rather than on the text. Third, the terms in the conventional AQE approaches are described as terms, while the terms in the WE-based approaches are described by real-number vectors with single dimension. Fourth, the conventional AQE approaches use the main dataset corpus, whereas the WE-based approaches use the main dataset corpus in addition to the WE corpus they created during the training task.

3. Conclusion

The conventional AQE approaches mainly rely on the assumption that each query term can select the best candidate terms based on their semantic closeness. The query semantics is analyzed locally, as prospect terms are chosen based on a one-word-at-a-time basis. However, this assumption is unable to represent the semantic of the query terms concerning the whole content of the query sentence.

The WE is the group name for a series of language modelling and functional learning techniques in NLP, where terms or phrases from vocabulary are described by real-number vectors. It comprises a mathematical

incorporation from a space with a single dimension per term to a continuous vector space with a significantly lower dimension (Diaz et al.,2016). The embedding models are normally trained in a broad corpus based on term proximity. For instance, the goal of the Word2Vec model is to predict the next word(s), i.e., the context window around the target word. This course aims to capture semantic and syntactic similarity between terms, since similar words often share similar contexts. The primary objective of many IR approaches is to model relevance (Saracevic,2016; Lavrenko & Croft,2017). In conclusion, the WE approaches seem to be more promising way for AQE than the conventional approaches.

References

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2018). Arabic query expansion using wordnet and association rules. In Information retrieval and management: Concepts, methodologies, tools, and applications (pp. 1239-1254). IGI Global.

Al-Chalabi, H., Ray, S., & Shaalan, K. (2015, April). Semantic based query expansion for Arabic question answering systems. In 2015 First International Conference on Arabic Computational Linguistics (ACLing) (pp. 127-132). IEEE.

ALMasri, M., Berrut, C., & Chevallet, J. P. (2016, March). A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In European conference on information retrieval (pp. 709-715). Springer, Cham.

Atwan, J., Mohd, M., Rashaideh, H., & Kanaan, G. (2016). Semantically enhanced pseudo relevance feedback for arabic information retrieval. Journal of Information Science, 42(2), 246-260.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR), 44(1), 1-50.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE), 2493-2537.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of documentation.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002, May). Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web (pp. 325-332).

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 1 pp. 367–377.

El, B. O. (2020). Document classification in information retrieval system based on neutrosophic sets. Filomat 34(1): 89-97.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2019). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. Journal of information science, 45(4), 429-442.

Farhan, Y. H., Mohd, M., & Noah, S. A. M. (2020). Survey of Automatic Query Expansion for Arabic Text Retrieval. Journal of Information Science Theory and Practice, 8(4), 67-86.

Farrar, D., & Hayes, J. H. (2019, May). A comparison of stemming techniques in tracing. In 2019 IEEE/ACM 10th International Symposium on Software and Systems Traceability (SST) (pp. 37-44). IEEE.

Fernández-Reyes, F. C., Hermosillo-Valadez, J., & Montes-y-Gómez, M. (2018). A prospect-guided global query expansion strategy using word embeddings. Information Processing & Management, 54(1), 1-13.

Hammo, B., Sleit, A., & El-Haj, M. (2007). Effectiveness of query expansion in searching the Holy Quran. The Second International Conference on Arabic Language Processing Rabat, Morocco. pp. 1-10.

Harris, Z. S. (1954). Distributional Structure. WORD 10(2-3): 146-162.

Jiang, J. J., & Conrath, D. W. (1996). A Concept-Based Approach To Retrieval From An Electronic Industrial Directory. International Journal of Electronic Commerce 1(1): 51-72.

Khafajeh, H., Yousef, N., & Kanaan, G. (2010, April). Automatic query expansion for Arabic text retrieval

based on association and similarity thesaurus. In Proceedings he European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS), Abu Dhabi, UAE.
Lavrenko, V. & Croft , W. B. (2017). Relevance-based language models. ACM SIGIR Forum. 51(2) pp. 260-267.
Lindén, K. & Piitulainen, J. (2004). Discovering Synonyms And Other Related Words. Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology. Geneva, Switzerland. pp. 63-70.
Mahgoub, A., Rashwan, M., Raafat, H., Zahran, M. & Fayek , M. (2014). Semantic Query Expansion For Arabic Information Retrieval. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). Doha, Qatar. pp. 87-92.
Mandala, R., Tokunaga, T., & Tanaka, H. (1998). The use of WordNet in information retrieval. In Usage of WordNet in Natural Language Processing Systems.
Maryamah, M., Arifin, A. Z., Sarno, R., & Morimoto, Y. (2019). Query expansion based on Wikipedia word embedding and BabelNet method for searching Arabic documents. International Journal of Intelligent Engineering & System, 12(5), 202-213.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ICLR: Proceeding of the International Conference on Learning Representations Workshop Track. Arizona, USA. abs/1301.3781 pp. 1301–3781.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193, 217-250.
Nwesri, A. F. A., & Alyagoubi, H. A. (2015, September). Applying Arabic stemming using query expansion. In 2015 26th international workshop on database and expert systems applications (DEXA) (pp. 299-303). IEEE.
Pal, D., Mitra, M., & Datta, K. (2014). Improving query expansion using WordNet. Journal of the Association for Information Science and Technology, 65(12), 2469-2478.
Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
Raza, M. A., Mokhtar, R., & Ahmad, N. (2019). A survey of statistical approaches for query expansion. Knowledge and information systems, 61(1), 1-25.
Raza, M. A., Mokhtar, R., Ahmad, N., Pasha, M., & Pasha, U. (2019). A taxonomy and survey of semantic approaches for query expansion. *IEEE Access*, *7*, 17823-17833.
Roy, D., Ganguly, D., Mitra, M., & Jones, G. J. (2016, October). Word vector compositionality based relevance feedback using kernel density estimation. In Proceedings of the 25th ACM international on conference on information and knowledge management (pp. 1281-1290).
Saracevic, T. (2016). The Notion Of Relevance In Information Science: Everybody Knows What Relevance Is. But, What Is It Really? Synthesis Lectures on Information Concepts, Retrieval, and Services 8(3): i-109.
Shaalan, K., Al-Sheikh, S., & Oroumchian, F. (2012, October). Query expansion based-on similarity of terms for improving Arabic information retrieval. In International conference on intelligent information processing (pp. 167-176). Springer, Berlin, Heidelberg.
Singh, J., & Sharan, A. (2017). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Computing and Applications, 28(9), 2557-2580.
Socher, R., Lin, C. C. Y., Ng, A. Y., & Manning, C. D. (2011, January). Parsing natural scenes and natural language with recursive neural networks. In ICML.
Vechtomova, O. (2009). Query Expansion For Information Retrieval. Dlm. Ö. M. T. e. LIU L. (pnyt.). Ed. Encyclopedia of database systems pp. 2254-2257. Springer.
Wang, X., Lai, G. & Liu, C. (2009). Recovering Relationships Between Documentation And Source Code Based On The Characteristics Of Software Engineering. Electronic Notes in Theoretical Computer Science 243: 121-137.
Wang, Y., Huang, H., & Feng, C. (2019). Query expansion with local conceptual word embeddings in

microblog retrieval. IEEE Transactions on Knowledge and Data Engineering.
White, R. W., & Horvitz, E. (2015). Belief dynamics and biases in web search. ACM Transactions on Information Systems (TOIS), 33(4), 1-46.
Z, H. (1968). Mathematical Structures Of Language. Number 21 In Interscience Tracts In Pure And Applied Mathematics. John Wiley and Sons 12: 13.
Zamani, H. & Croft , W. B. (2016). Embedding-Based Query Language Models. Proceedings of the 2016 ACM international conference on the theory of information retrieval. Newark Delaware USA. pp. 147-156.
Zamani, H. & Croft , W. B. (2017). Relevance-based word embedding. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 505-514.

Recommendation for Group of Users with LOD

Rosmamalmi Mat Nawi[a]*, Shahrul Azman Mohd Noah[b], Lailatul Qadri Zakaria[c]

[a,b,c] Centre of Artificial Intelligence Technology, Faculty of Technology and Information Science, The National University of Malaysia.
* Email: rosmamalmi@siswa.ukm.edu.my

Abstract

This paper presents a model-based movie recommender system by exploiting Linked of Data (LOD) technology to enrich the item information. While numerous studies have been conducted on both conventional recommender systems and Linked Data-enabled recommender systems, little effort has been made to investigate the potentials LOD in the context of group formation. Our proposed framework adopts the LOD-based group recommender system model before the formation of group to enhance the quality of recommendation by tackling the sparseness issue. The findings indicate that the proposed model leads to a better result compared to the baseline. The GRS-LOD model improves prediction accuracy, as indicated by the low RMSE and MAE errors, and exhibits good recommendation relevancy considering the high Precision, Recall, and F1-Score metrics.

Keywords: Group Recommendation System; Linked Open Data; sparsity

1. Introduction

Recommender Systems (RS) are information filtering tools in the decision-making process that recommend items to users. (El-Ashmawi et al., 2020). Although conventional research on RS has almost entirely focused on offering recommendations to single users, there are several other scenarios where the system needs to provide items to groups of users. Group Recommender Systems (GRS) has a significant challenge to provide recommendations if insufficient data is received from group member rating (Ghazarian et al., 2014) since it relies heavily on the available data sources. Furthermore, it is difficult to identify the most similar users when there are insufficient data items ratings (Khusro et al., 2016).

Linked Open Data (LOD) datasets provide additional data that can be integrated into several domains (e.g., film) to augment item information. Hence, LOD may be beneficial when such a number or quality of a ready dataset is insufficient. RS can effectively exploit LOD to deal with classical problems of cold-start and sparsity. Even though it has been widely implemented in conventional RS from many perspectives (references LOD untuk individual recommendation), the LOD technology approach in GRS is still underexplored. Thus, the research presented in this paper proposed the use of LOD for group recommendation. The proposed model aims to offer a method to tackle the sparseness issue prior to user group formation since limited rating affects clustering.

In a nutshell, the contributions of our work are as follows:

- We proposed a GRS-LOD model that applies the LOD technology in enriching the item information to be applied in GRS.
- We presented an approach for enhancing user information using similarity-based DBpedia attributes (*'dbo:director'* and *'dbo:starring'*).
- We illustrate how the model could be utilized to form a user cluster more efficiently by dealing with data sparsity in group member profiles.

2. Proposed Methodology

Our proposed method exploited the LOD information to enrich the movie information. Thus, it may further enhance the effectiveness of the clustering process. We selected two persistent attributes and representative of the movie domain in which we performed the evaluation. Further resources could be iteratively expanded based on these attributes. Movies in DBpedia, for example, provide essential details such as star cast and director. As illustrated in Fig. 1, additional information about the actor who starred in the movie can be explored through the LOD (e.g., the relation *'dbo:starring'* existing between *'Keanu Reeves'* and *'The Matrix'*).

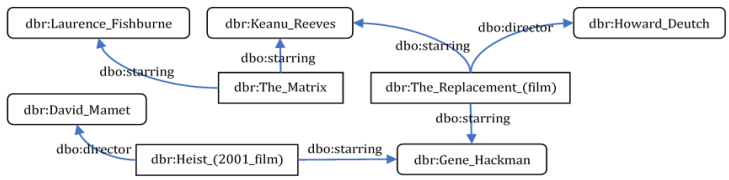


Fig. 1. Movie relation based on DBpedia attribute

The research framework employed in this study is as illustrated in Fig. 2. It differentiates the workflow applies between baseline and GRS-LOD model with four main components.

The GRS-LOD model is built on the first component, which comprises five major processes. The first two processes, 'ML1M-DBpedia mapping' and 'DBpedia data extraction', include linking and extracting the two datasets. While the third process entails data filtering and integration once the data has been enriched with DBpedia information. A pre-clustering phase is employed in the fourth process, 'On-attributes similarity', that finds similarities between users based on the investigated attributes. The generated user cluster based on the attributes is then subjected to a rating prediction based on attribute similarity. Note that we implement the rating prediction for five users on each selected data of attribute.

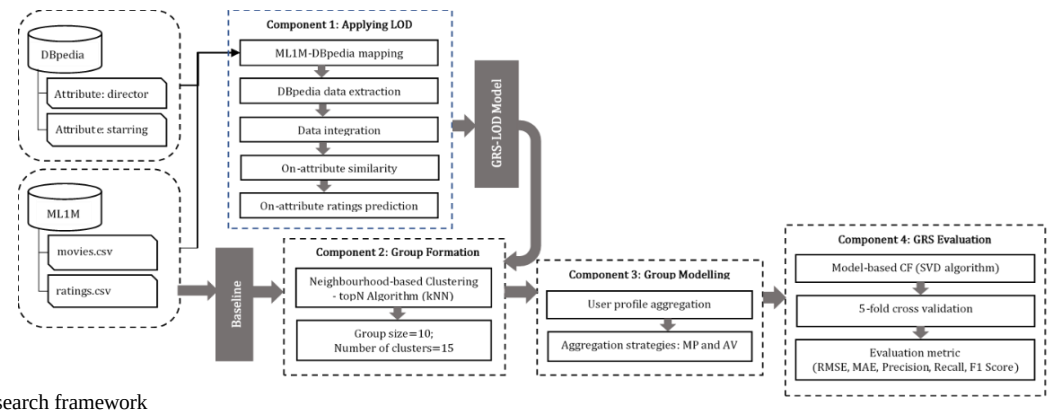


Fig. 2. Research framework

The GRS-LOD model produces an additional rating dataset based on the DBpedia attributes. The model is then used to build clusters using the k-Nearest Neighbour (kNN) algorithm. This approach clusters homogeneous user with an automatically detected group, and it alludes to the second component. The basic principle of neighbourhood-based clustering is to find similarities between users. It represents each user's neighbourhood is those other users who are most similar to him. We assume that two people have comparable interests and are similar if they rated the movie similarly. We use cosine similarity in this study.

We apply the Average (AV) *(1)* and Most Pleasure (MP) *(2)* aggregation strategies along with the profile aggregation approach. A brief description of each strategy, which $Grel(G,i)$ represents the group preferences for the item i, Rel_{ui} is the user preference u for the item i, and the group preferences is represented by G.

$$Grel(G,i) = \frac{\sum_{u \in G} Rel_{ui}}{|G|} \quad (1)$$

$$Grel(G,i) = max_{u \in G}(Rel_{ui}) \quad (2)$$

We employ the model-based CF method, namely the SVD algorithm, to generate the recommendations for the groups. Recommendation through GRSs represented as in *(3)* with G reference to the target group, I is a set of available items $Predicition(G, i_k)$ and are utility functions for items i_k based on group members G.

$$Recommendation\ (G,I) = arg\frac{max}{i_k \in I}\ Prediction(G, i_k) \quad (3)$$

3. Result and Discussion

As for experimental evaluation, we used the Movielens 1 Million (ML1M) dataset, which comprises 6040 users and 3952 movies with a total of 1,000,209 ratings. DBpedia's two attributes had been chosen: *'dbo:director'* and *'dbo:starring'*. We employed 5-fold cross-validation over 15 formed groups and evaluated the GRS-LOD model's effectiveness using a consistent ten-size group (ten users per cluster).

According to Table 1, the evaluation score for the GRS-LOD model outscored the baseline findings for both the MP and AV strategies. It appears that the GRS-LOD model is capable of improving prediction accuracy (Figure 3 (a)) as evidenced by the low RMSE and MAE errors, and also shows good recommendation relevancy (Figure 3 (b)) based on the high performance of the Precision, Recall and F1-Score metric. In terms of aggregation strategies, compared to MP, AV strategies performed better either as illustrated in Figure 3(a) and 3(b).

Table 1. Evaluation Score of Baseline and GRS-LOD Model

	Baseline		GRS-LOD Model	
	MP	AV	MP	AV
Mean RMSE	1.0246	0.9178	1.0141	0.8983
Mean MAE	0.8266	0.7117	0.8108	0.6881
Precision (*k*=5, *threshold*=3.5)	0.9200	0.9333	0.9467	0.9600
Recall	0.0497	0.0515	0.0576	0.0621
F1 Score	0.0943	0.0976	0.1085	0.1167

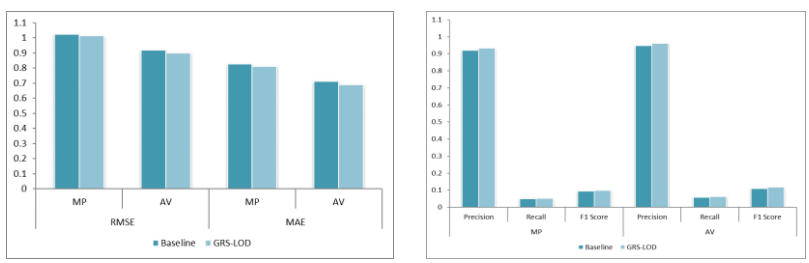


Fig. 3. (a) Prediction Accuracy Graph; (b) Recommendation Relevancy Graph

These findings demonstrate that the proposed method facilitates effective group clustering implementation and emphasizes that additional rating data allows for more effective group recommendations since more data can identify user similarity. It underlines that the more similar the users in a group are, the more valuable the group's recommendation is since the sparsity being reduced. As stated by Wang et al. (2016), generating effective recommendations is more onerous if a large number of data is not rated. Thus, highlighting the data insufficiency in the group profile is a significant practice in delivering quality and relevant recommendations

for groups. Therefore, it is advantageous to cluster the user for a group based on their features. The more similar the user preferences are in the group, the better the group recommendations (Pessemier et al., 2014).

4. Conclusion

This work discusses the issue of sparsity in the context of GRS and applies the proposed model, GRS-LOD, in the group formation process applying LOD technology. The acquired findings demonstrate the richness of the data retrieved from DBpedia can enhance group formation. The experimental findings further prove that the proposed method may generate acceptable and quality group recommendations. Thus, we believe this work represents a preliminary step towards a new generation of LOD-enabled GRS as it can be explored to other new dimensions in GRS. In future work, we will investigate the context of explaining the recommendation for groups since it facilitates group members to grasp the system's recommended items.

Acknowledgements

The authors gratefully acknowledge the sponsorship received to carry out this study from Tun Hussein Onn University of Malaysia and from the Malaysia Ministry of Higher Education.

References

Ahmed, E. Ben, Tebourski, W., Karaa, W. B. A., & Gargouri, F. (2015). SMART : Semantic Multidimensional Group Recommendations. *Multimedia Tools and Applications*, *74*, 10419–10437.
Alam, M., & Biswas, R. (2019). *Linked Open Data Validity - A Technical Report from ISWS 2018*.
El-Ashmawi, W. H., Ali, A. F., & Slowik, A. (2020). Hybrid crow search and uniform crossover algorithm-based clustering for top-N recommendation system. *Neural Computing and Applications*, *6*.
Garcia, I., Pajares, S., Sebastia, L., & Onaindia, E. (2011). Preference elicitation techniques for group recommender systems. *Information Sciences*, *189*, 155–175. https://doi.org/10.1016/j.ins.2011.11.037
Ghazarian, S., Shabib, N., & Nematbakhsh, M. A. (2014). Improving Sparsity Problem in Group Recommendation. *Proceedings of the 25th ACM Hypertext and Social Media Conference (Hypertext 2014), Santiago, Chile, September 3*.
Khusro, S., Ali, Z., & Ullah, I. (2016). Recommender Systems : Issues, Challenges, and Research Opportunities. *Information Science and Applications (ICISA)*, 1179–1189.
Masthoff, J. (2015). Group Recommender Systems: Aggregation, Satisfaction and Group Attributes. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook, Second Edition* (pp. 743–776). Springer Science Business Media New York. https://doi.org/10.1007/978-1-4899-7637-6_22
Nawi, R. M., Noah, S. A. M., & Zakaria, L. Q. (2020). Evaluation of Group Modelling Strategy in Model-Based Collaborative Filtering Recommendation. *International Journal of Machine Learning and Computing (IJMLC)*, *10*(2).
Pessemier, T. De, Dooms, S., & Martens, L. (2014). Comparison of Group Recommendation Algorithms. *Multimedia Tools Application*, *72*(3), 2497–2541.
Roy, S. B., Lakshmanan, L. V. S., & Liu, R. (2015). From Group Recommendations to Group Formation. *SIGMOD '15: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1603–1616.
Wang, W., Zhang, G., & Lu, J. (2016). Member contribution-based group recommender system. *Decision Support Systems*, *87*, 80–93. https://doi.org/https://doi.org/10.1016/j.dss.2016.05.002

Review of Malay Named Entity Recognition

Hafsah[a*], Saidah Saad[b], Lailatul Qadri Zakaria[c]

[a,b,c] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia
*Email: p93826@siswa.ukm.edu.my

Abstract

Named Entity Recognition (NER) is a technique used extensively to extract useful information from unstructured natural language document collections. Named-Entity Recognition has important information extraction tasks that should be developed for all languages in the world and almost all domains. Most of the research on NER has been done for English languages. The Malay language NER cannot use the English corpus because there are differences in speech structure and morphology between English and Malay. Based on discussion shows that the research of Malay named entity still in the early stage.

Keywords: Named Entity Recognition; Natural Language Processing; Malay language; NER approach

1. Introduction

Named-Entity Recognition (NER) is a sub-part of Natural Language Processing (NLP) research which is included in the field of Artificial Intelligence (AI). Named Entity Recognition (NER) is the initial step in information extraction that seeks to find and classify entities mentioned in the text into predetermined categories, such as the name of the person, organization, location, expression, time, amount, monetary value, percentage, etc. (Saad & Mansor, 2018).

Currently, research related to NER has been carried out for various purposes and the methods used. The methods used also vary, from rule-based to the use of Machine Learning (ML) (Saad & Mansor, 2018). The rule-based approach uses defined rules based on linguistic knowledge with analysis carried out at the syntactic and semantic levels (Goyal et al., 2018). This method has limitations because we have to define as many rules as possible to get optimal results (Nadia & Omar, 2019). To overcome these limitations, we can use the ML approach to study patterns from the data by only providing sufficient data sets (Salini et al., 2017).

An approach that is also widely used recently is to use deep learning (DL) to recognize patterns of entities in sentences (Li et al., 2020). Named-Entity Recognition has important information extraction tasks that should be developed for all languages in the world and almost all domains. However, these tasks differ according to language, domain, and systems development approach (Patil et al., 2019).

Most of the Named Entity Recognition research focuses on English as well as European languages. But along with the development of research in this field, more and more types of languages have been researched. English and Japanese are well explored in MUC-6 [5] and earlier works. German, Dutch and Spanish is discussed at the CONLL conference. Chinese is studied in an abundant literary language as well as French, Greek and Italian. Arabic has started to receive a lot of attention in large-scale projects such as Global Autonomous Language Exploitation (GALE). In time, Asian and several other languages were also considered (Goyal et al., 2018).

This paper is organized as follows. The second section of this paper presents the Literature Review of NER. This section will discuss the approaches that the previous researcher has done. The third section gives an overview of issues and challenges in Malay Named entity recognition.

2. Literature Review

Named entity recognition has three approaches classified into three main streams: rule-based, machine learning approaches, and hybrid approaches.

2.1. Rule-Based Named Entity Recognizers

Rule-based and dictionary-based methods are the earliest methods used in NER (Ji et al., 2019). They rely on handcrafted rules, use named entity libraries, and assign weights to each rule. When a rule conflict is encountered, the rule with the highest value is selected to determine the named entity type. However, these rules often depend on the specific language, domain, and text style (Ji et al., 2019). Alfred et al. in 2014 have proposed the Malay-named entity using Malay articles. His approach is based on a rule-based part of speech (POS) tagging process and contextual feature rules. Some dictionaries are also manually created to detect three named entities: a person, location, and organization. Evaluation using standard performance metrics has shown where Recall of 94.44%, Precision of 85%, and F-score of 89.47%.

Wulandari et al. (2018) have conducted research related to NER on biological documents using rule-based and naïve Bayes classifier methods. This study used 19 training documents. The document was processed and annotated manually based on NEs and obtained 1,135 training data in the form of words. The pre-processing of data includes POS-tagging and n-gram. From the combination of rule-based and naïve Bayes methods, this study obtained an average Precision, Recall, and F-measure of 0.8 with a micro average.

In Malay, research related to criminal news documents was conducted by Saad and Mansor (2018). This research builds a crime news corpus sourced from BERNAMA news. Linguists manually check the corpus to identify name entities such as individuals, organizations, locations, dates, times, finances, percentages, crimes, and weapons. This prototype system's testing shows good promising results with a Recall value of 78.67%, Precision 71.11%, and F-measure 74.7%.

Nadia and Omar (2019) proposed Malay NER Using Rules-Based. This Research Identifies Name Entities Involving Nine Name Entities: Individual Name, Location, Organization, Position, Date, Time, Finance, Measurement, And Percentage. This test shows promising results with a Recall value of 92.13%, a Precision value of 90.23%, and an F-Score of 91.05%.

2.2. Machine Learning-Based Named Entity Recognizers

The Machine Learning method is used to classify and uses a statistical classification model to recognize named entities. This method looks for patterns and their relationship in a text, tries to create models with a statistical approach and machine learning algorithms, and identifies and classifies nouns into several classes, such as a person, location, and time (Jurafsky and Martin, 2017). Surwaningsih et al. (2014) conducted a study on Indonesian Medical Named Entity Recognition (ImNER) utilizing a Support Vector Machine (SVM). They used data in the form of 3,000 sentences taken randomly. The accuracy value obtained is 90%. Their research uses data on word types, contextual characteristics of words, word writing systems, and common word lists. Apart from that, they also make use of medical-related word lists.

Aryoyudanta et al. (2016) use the Co-Training algorithm to empower unlabeled data to obtain new labeled data. This study uses news articles as unlabeled data and Dbpedia as labeled data. This research's initial stage is to perform text-processing on unlabeled data, which POS-Tagging then follows. The purpose of POS-Tagging is to look for words that are most likely to have named entities. Furthermore, they use the Co-Training algorithm

to do entity labeling on unlabeled data with data from DBpedia. For testing, they use the SVM algorithm for labeled data modeling. The results obtained in this study are a precision value of 73.6%, a recall value of 80.1%, and an F1 of 76.5%.

Bhasuran et al. (2016) have proposed a biomedical NER based on a stacked ensemble approach. The authors applied several domain-specific, morphological, orthographical, and contextual features, Conditional Random Fields (CRF) based modeling, and two fuzzy matching algorithms for extracting disease-named entities. Some post-processing measures are also applied to enhance the performance of the model.

Salleh et al. (2017) propose that the Malay language NER uses the Python CRFsuite and several features. The feature such as capitalization, lowercase, previous and closest words, digits, word forms, and word POS tags, and others show the potential for increasing the accuracy of results from recognizing named entities Malay. Salleh et al. (2018) proposed Malay NER using the fuzzy c-means method with the Rapid Miner software and dataset from Bernama Malay news. The types of named entities analyzed are person, location, organization, and facility. In conclusion, the overall percentage accuracy gave markedly good results based on clustering matching with 98.57%.

2.3. Hybrid Named Entity Recognizers

A hybrid Named Entity recognition system combines both rule-based and machine learning techniques. These new methods combine the strongest points from each method: the adaptability and flexibility from machine learning approaches and rules to improve efficiency. Keretna et al. (2014) present a hybrid model comprising the rule-based and lexicon-based techniques for extracting drug Named entity from the informal and unstructured medical text. The experimental outcome indicates that integrating many valuable rules into a lexicon-based technique can enhance the performance of the BioNER problem. The proposed model can achieve an f-score of 66.97%.

Munkhjargal et al. (2015) have introduced a Mongolian named entity recognizer. The authors used statistical techniques, namely Maximum Entropy, SVM, CRF, gazetteers, and string matching patterns, to handle the vocabulary words. The optimal ensemble reached 90.59% precision, 85.88% recall, and 88.17% F1 score.

3. Issues and Challenges in Malay Named Entity Recognition

Most of the documents on a website are unstructured, making it difficult to get the relevant information in structured data. Information extraction is the process of converting unstructured data into structured data. Thus the extraction of named entities is a challenging task. Apart from the techniques used, several factors affect NER tasks' performance, such as language factors, domain factors, entity type factors, etc. Several researchers have researched the Malay language NER. Most of the research of NER in Malay uses a Rule-based approach and a supervised system approach (Nadia & Omar, 2019).

The NER system's performance is highly dependent on some language resources such as POS tagger, morphological analyzer, chunker, parser, etc. The Malay language has some similarities with English features such as capitalization and word POS tag such as proper noun to recognize the entity (Morsidi et al., 2016). The supervised named entity recognition system requires large annotated corporations to classify named entities from the test data. The challenge because the Malay language corpus is still limited compared to the English corpus. The Malay language NER cannot use the English corpus because there are differences in speech structure and morphology between English and Malay (Nadia & Omar, 2019).

Domain factors have a significant influence on the Named Entity Recognition task. Various domains are explored for NER assignments, such as news articles, crime, medical, etc.

4. Conclusion and Future Work

Named Entity Recognition is an area of research steadily increasing due to its significant contribution to many natural language applications. NER has a vital role to play in automated information extraction. In this article, NER approaches from rules-based to machine learning and hybrid approaches are presented. Also, challenges and issues related to NER have been outlined. Based on discussion shows that the research of Malay named entity still in the early stage. The demand for Malay NER will continue to expand and need a lot of consideration to matured it for the other application development based on Malay text and document to be in proper place compare to other languages.

Acknowledgements

This research was supported by Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia and this project is funded by MoHE under research code FRGS/1/2019/ICT02/UKM/02/2.

References

Alfred, R., Leong, L. C., On, C. K. & Anthony, P. (2014). Malay Named Entity Recognition Based on Rule-Based Approach. International Journal of Machine Learning and Computing, 4(3), 300-306.

A Goyal, M Kumar, and V Gupta (2018). Recent named entity recognition and classification techniques: a systematic review. Computer Science Review 29, 21-43

Aryotudanta, B., Adji, T.B. & Hidayah, I. (2016). Semi-Supervised Learning Approach for Indonesian Named Entity Recognition (NER) using Co-Training Algorithm. 2016 International Seminar on Intelligent Technology and Its Application, 7-11.

Bhasuran, B., Murugesan, G., Abdulkadhar, S. & Natarajan, J. (2016). Stacked Ensemble Combined with Fuzzy Matching for Biomedical Named Entity Recognition of Diseases, Journal of Biomedical Informatics, 64, 1-9.

D. W. Wulandari, P. P. Adikara, & S. Adinugroho. (2018). Named Entity Recognition (NER) pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(11), 4555-4563.

Jing Li, Aixin Sun, Jianglei Han, & Chenliang Li (2020). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions On Knowledge And Data Engineering.

Jurafsky, D., & Martin, J. H. (2017). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3nd., Vol. 21). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Keretna, S., Lim, C. P., & Creighton, D. (2014, June). A hybrid model for named entity recognition using unstructured medical text. In 2014 9th International Conference on System of Systems Engineering (SOSE) (pp. 85-90). IEEE.

Mohd Noor, N., Sulaiman, J., & Noah, S. A. (2016). Malay Name Entity Recognition Using Limited Resources. Advanced Science Letters, 22(10), 2968–2971.

Morsidi, F., Sarkawi, S., Sulaiman, S., Mohammad, S. A., & Wahid, R. A. (2015). Malay named entity recognition: a review. Journal of ICT in Education, 2, 1-14.

Munkhjargal, Z., Bella, G., Chagnaa, A., & Giunchiglia, F. (2015, September). Named entity recognition for Mongolian language. In International Conference on Text, Speech, and Dialogue (pp. 243-251). Springer, Cham.

Nadia, U. and Omar, N. (2019). Malay Named Entity Recognition Using Rule-Based Approach, Jurnal Teknologi Maklumat Dan Multimedia Asia-Pasifik, 8(1), 37 – 47.

Patil, N., Patil, A., & Pawar, B. V. (2020). Named entity recognition using conditional random fields. Procedia Computer Science, 167, 1181-1188.
Saad, S., & Mansor, M. K. (2018). Named entity recognition approach for Malay crime news retrieval. GEMA Online Journal of Language Studies, 18(4), 216-235.
Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2017, May). A Malay Named Entity Recognition using conditional random fields. In 2017 5th International Conference on Information and Communication Technology (ICoIC7) (pp. 1-6), IEEE.
Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2018). Named Entity Recognition using Fuzzy C-Means Clustering Method for Malay Textual Data Analysis. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(2-7), 121-126.
Suwarningsih, W., Supriana, I., & Purwarianti, A. (2014, August). ImNER Indonesian medical named entity recognition. In 2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment (pp. 184-188). IEEE.
Salini, U. J. A., & Jeyapriya, U. (2017). Named Entity Recognition Using Machine Learning Approaches. Int. J. Innov. Res. Sci. Eng. Technol, 6(11), 491-501.
Ji, Y., Tong, C., Liang, J., Yang, X., Zhao, Z., & Wang, X. (2019, February). A deep learning method for named entity recognition in bidding document. In Journal of Physics: Conference Series (Vol. 1168, No. 3, p. 032076). IOP Publishing.

Significance of Player Elimination in Battle Royale Games Popularity

Sagguneswaraan Thavamuni[*a], Muhammad Nazhif Rizani [b], Mohd Nor Akmal Khalid [c], Hiroyuki Iida [d]

Japan Advanced Institute of Science and Technology, School of Information Science, Nomi, Ishikawa, Japan 923-1211
* Email: saggu[a],nazhif[b], akmal[c], iida[d]@jaist.ac.jp

Abstract

Battle royale games have been on the rise of popularity with over 350 million unique players and total revenue of over $1 billion since their first major breakout in 2017 with PlayerUnknown's Battlegrounds (PUBG) and Fortnite. Battle royale games often housed many players where players are gradually eliminated over time until there is only one player remaining. The popularity of battle royale games are shown in the sheer number of players and the revenue from the sales of the game or in-game purchases such as skins. This paper aims to analyze the game progress using motion in mind theory to find the similarities and differences between three battle royale games: PUBG, Fortnite, and Fall Guys, and explain the popularity of the games focusing on the reducing number of players over time due to player eliminations. The mechanism of starting with high number of players and slowly eliminating them might be the reason why battle royale games are interesting and have a high player retention. It is expected such games are challenging and engaging enough for players to ensure that they will stick with the game for a longer time.

Keywords: Battle Royale, Player Elimination, Game Refinement, Motion in mind

1. Introduction

Battle Royale is a game genre where multiple players are loaded to the game, and players are constantly being eliminated until the last remaining player, which will be crowned the winner. This genre was inspired by the Japanese film titled Battle Royal, which features a last-man-standing competition (Rahkar Farshi, 2020). The battle royale genre rose in popularity in 2016 with the launch of PlayerUnknown's Battlegrounds (PUBG) and Fortnite. Both PUBG and Fortnite maintained their top spot as the most popular battle royale game, having over 350 million unique players and total revenue of over $1 billion (Iqbal, 2021). Both PUBG and Fortnite are also available on various platforms such as PC, PlayStation, Xbox, and mobile, allowing them to reach a wider audience (Al-Mansour, 2019). Fortnite and PUBG mobile have a free-to-play business model with sales of skins, battle passes, and other items in their in-game shop. This freemium model makes it more accessible to children, hence explaining these games' popularity among kids (Carter et al., 2020).

Game refinement theory is used to explain the balance between skill and chance, focusing on the uncertainty of the game's outcome (Iida & Khalid, 2020). A good balance of skill and chance is essential to ensure the game is competitive and fun. Based on the game refinement theory, the motions in mind are studied further to explain our minds' motion. Both theories will analyze the battle royale genre and see why these games are interesting even though they might seem repetitive. This paper will focus on three popular Battle Royale genre games: PUBG, Fortnite, and Fall Guys.

2. Motion in Mind in Battle Royale Games

Game refinement theory is derived from the game progress model which is focused on the game speed or scoring speed and is used to evaluate the $\boldsymbol{GR}$ measure for the game. $\boldsymbol{GR}$ measure of various boardgames and sports have been evaluated previously which shows that most sophisticated game has a $\boldsymbol{GR}$ measure of between 0.07~0.08, which is called the $\boldsymbol{GR}$ zone (Sutiono et al., 2014). Games in this zone are sophisticated games where players are engaged and excited for the outcome of the game as its unpredictable. As acceleration in games are measured as $\boldsymbol{GR}^2$, we can now evaluate other physics values as well. Hence, it is estimated that the velocity $\boldsymbol{v}$ can be evaluated as the total number of goals G divided by the total shots attempts T. Fig 1 show some of the formula used to analyze the games in this paper (Iida & Khalid, 2020). The key points are that the mass $\boldsymbol{m}$ signifies the challenge or difficulty of the game while the momentum $\overrightarrow{\boldsymbol{p}_1}$ shows the players engagement.

$$\boldsymbol{v} = \frac{B}{2D} \; or \; \frac{G}{T}$$

where B: branching factor, D: depth, G: Goals, T: Total Attempts at Goal

$$\boldsymbol{m} + \boldsymbol{v} = 1$$

$$\overrightarrow{\boldsymbol{p}_1} = mv$$

$$\boldsymbol{E_p} = 2mv^2$$

$$\boldsymbol{E_p} = \overrightarrow{p_1} + \overrightarrow{p_2}$$

Fig. 1 List of formulas used in motion in mind theory

Both PUBG and Fortnite is a third-person shooter with up to 100 players at a time, where the goal of the game is to be the last person standing. Players are loaded on a flying vehicle at the start of the game, and they will have to choose a place for them to land to start the game. Once they land, they will have to scavenge the area to find guns and various other items to help them survive. As the map is quite huge, to ensure that players will engage with each other, the play area will gradually shrink throughout the game's duration. This situation is done using a game concept called "The Circle" that shows the new play area (also known as the safe area), and players will need to move to the new area as soon as possible. Any players outside the safe area will gradually take damage and be eliminated if they are there for too long, ensuring that players will meet other players as they move to the new area, speed up the game, and leave only one survivor at the end. The circle shrinks in fixed time intervals throughout the game, but the new circle's location will be randomized within the current circle. For Fortnite, there is a possibility that the new circle might be outside of the current safe area, forcing players to take some damage to move to the new safe area. Also, players can loot resources like wood and metal in Fortnite, which they can use to build structures such as ladders, walls, and roofs to help provide some offensive or defensive edge.

Unlike PUBG and Fortnite, Fall Guys is not a third-person shooter game but is a more fun take on the Battle Royale genre. Fall Guys is a multiplayer game where the players must navigate through mini-games and obstacles, like the obstacles in famous shows like "Takeshi's Castle" and "Total Wipeout" (Gene Park, 2020). The game starts with 60 players and gradually reduced until only one player is remaining. Players who do not make it to the finish line of the obstacles in time or those that got the last place in the mini game will be eliminated. The game lasts about 4 to 6 rounds, and each round has its obstacle course selected randomly.

3. Findings and Analysis

To analyze these three games, we will need to formulate the velocity $\boldsymbol{v}$ of each game. As both PUBG and Fortnite are similar styled game, we can formulate the $\boldsymbol{v}$ based on the formula for from Fig 1, as shown in Eq. 1 below. As the play area shrinks in fixed time intervals, we can assume each time the circle shrinks as a round.

For Fall Guys, as of the time of writing, matches data are not available to the public. Hence, we collected data from a player who revealed their play data based on 500 matches that they played (Monkmanny, 2020). As Fall Guys players are eliminated based on the obstacles each round, we used the total rounds qualified to calculate the velocity. Eq. 1 shows the formula of $\boldsymbol{v}$ for the case of Fall Guys.

$$v_{PUBG/Fortnite} = \frac{Total\ Kills\ in\ Round}{Total\ Players\ in\ Round} \qquad v_{Fall\ Guys} = \frac{Total\ Rounds\ Qualified}{Total\ Rounds} \tag{1}$$

Based on and $\boldsymbol{GR}$ measure and the velocity, mass, momentum, and potential energy using the formulas in Fig 1, the data for PUBG, Fortnite and Fall Guys are given in Table 1 and illustrated in Figure 2.

Table 1. Mass, velocity, and GR measure for PUBG, Fortnite and Fall Guys across rounds

	PUBG			**Fortnite**			**Fall Guys**		
Rounds	$\boldsymbol{m}$	$\boldsymbol{v}$	$\boldsymbol{GR}$	$\boldsymbol{m}$	$\boldsymbol{v}$	$\boldsymbol{GR}$	$\boldsymbol{m}$	$\boldsymbol{v}$	$\boldsymbol{GR}$
1	0.880	0.120	0.009	0.927	0.073	0.008	0.010	0.990	0.044
2	0.768	0.232	0.014	0.935	0.065	0.008	0.109	0.891	0.042
3	0.621	0.379	0.020	0.961	0.039	0.006	0.136	0.864	0.044
4	0.615	0.385	0.026	0.872	0.128	0.012	0.307	0.693	0.043
5	0.550	0.450	0.036	0.959	0.041	0.007	0.450	0.550	0.049
6	0.612	0.388	0.046	0.925	0.075	0.010	0.875	0.125	0.125
7	0.408	0.592	0.078	0.901	0.099	0.012	-	-	-
8	-	-	-	0.913	0.087	0.012	-	-	-
9	-	-	-	0.957	0.043	0.009	-	-	-
10	-	-	-	0.020	0.980	0.043	-	-	-

From Table 1, we can see that the value gradually increases and approaches the zone value of 0.07 as the game progresses for all three analyzed games. This condition shows that the game gets a good balance of luck and skill at the later stages of the games. In a previous paper, we conjectured that games get more exciting and stochastic when the number of players increases in the game, and conversely, the game will be more deterministic if it has lesser players (Thavamuni et al., 2018). However, from our findings here, we can see that this is not the case. The values of all three games increase as there are lesser players at the game's ending phases. As each stage reduces the number of players in the game, this shows that the number of players of battle royale games is indirectly proportional to the games' stochasticity.

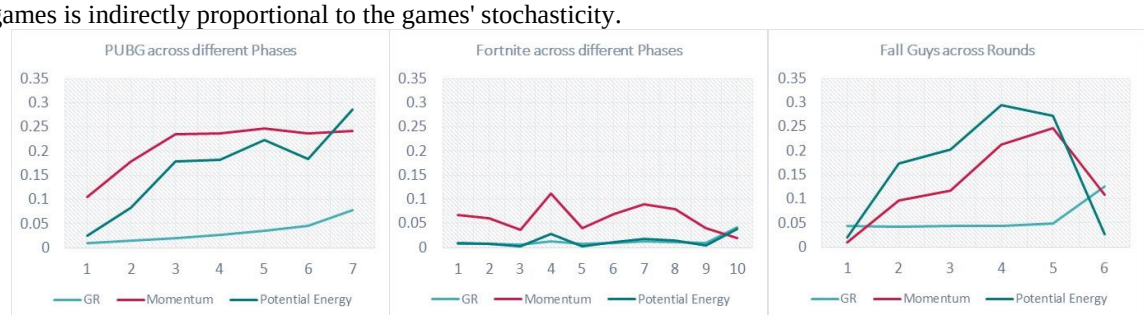


Fig. 2. Graphs of GR, momentum and potential energy across rounds/phased for PUBG, Fortnite and Fall Guys

The three games have different patterns of mass in mind. In PUBG and Fortnite, the pattern gradually decreases and fluctuates slightly but drastically drops in the final round, respectively. Conversely, our findings

from Fall Guys show that the gradually increases across the rounds. In terms of the game's challenge and difficulty, PUBG and Fortnite become less difficult throughout the game while Fall Guys become more and more challenging. The constant shift in difficulty in the games causes the player to enjoy the game. Although PUBG and Fortnite are difficult initially, it is still very popular to kids despite the challenge. Therefore, the fluctuation makes the game popular to both players and viewers even after years of the game's launch. The momentum also showed an increasing trend throughout the game and decreased from Fortnite having a sharp increase and decrease of momentum at phase 4. This condition shows that the player's engagement increases as they play the game despite reducing the difficulty in PUBG and Fortnite. It explains why players get addicted to battle royale games despite it being difficult to win as they would like to achieve a higher stage each time they play, increasing their retention of the game.

4. Conclusion

The popularity of battle royale games maintained its popularity throughout the test of time and remains one of the most profitable game genres in 2021, with a few more titles will be launched this year (Christopher Livingston, 2021). This genre's popularity was justified from this study where all three games increase throughout the game and approach the zone value of 0.07 at the ending stages of the game. The change in difficulty as the game progresses helps to make players engaged throughout the game and come back for another round when they are done. The momentum also shows a general increase throughout the games, making it a very engaging experience for players. Hence, the games become more exciting and engaging to the players as they survive through each round, giving them the thrill of achieving a higher position. In the future, it would be interesting to compare with a variety of newer battle royale games to see if the trend consistent and analyze the e-sports capability of this genre compared to other popular e-sports such as Dota 2, League of Legends, and Counter-Strike.

References

Al-Mansour, J. (2019). The success behind the PuBG era: A case study perspective. Academy of Strategic Management Journal, 18(6), 1–17.

Carter, M., Moore, K., Mavoa, J., Horst, H., & Gaspard, L. (2020). Situating the Appeal of Fortnite Within Children's Changing Play Cultures. Games and Culture, 15(4), 453–471.

Christopher Livingston, M. P. (2021). The Best Battle Royale Games of 2021. https://www.pcgamer.com/battle-royale-games/

Gene Park, E. F. (2020). Fall Guys: Ideas to improve an already fun game - The Washington Post. https://www.washingtonpost.com/video-games/2020/08/12/fall-guys-ideas/

Iida, H., & Khalid, M. N. A. (2020). Using games to study law of motions in mind. IEEE Access, 8, 138701–138709.

Iqbal, M. (2021). Fortnite Usage and Revenue Statistics 2020 - Business of Apps. https://www.businessofapps.com/data/fortnite-statistics/

Monkmanny. (2020). I recorded data on 100 matches for Fall Guys. https://www.reddit.com/r/FallGuysGame/comments/i9xq1g/

Rahkar Farshi, T. (2020). Battle royale optimization algorithm. Neural Computing and Applications, 1–19.

Sutiono, A. P., Purwarianti, A., & Iida, H. (2014). A mathematical model of game refinement. International Conference on Intelligent Technologies for Interactive Entertainment, 148–151.

Thavamuni, S., Ismail, H., & Iida, H. (2018). Analysis of the Effect of Number of Players on the Excitement of the Game with Respect to Fairness. International Conference on Entertainment Computing, 139–151.

Entertainment Analysis of Animation Based on GR Theory

Wang Xinyue[a]*, Mohd Nor Akmal Khalid [b], Hiroyuki Iida[c]

[a]School of Information Science, Japan Advanced Institue of Science and Technology, 923-1211 Ishikawa, Japan
[b]Research Center for Entertainment Science, Japan Advanced Institue of Science and Technology, 923-1211 Ishikawa, Japan
* Email: s1910403@jaist.ac.jp

Abstract

Game refinement (GR) theory has been applied to many different games and the entertainment impact in games has been quantified. The purpose of entertainment is to make people feel happy and satisfied, but not superficially flatter the audience's emotional needs. Animation, as a form of artistic expression that combines many artistic disciplines such as painting, film, digital media, photography, music, and literature, was initially treated as an art form to entertain children, but nowadays it is also entertaining adults. And excellent animation is not only artistic, but also entertaining. In this paper, the entertainment impact in animation through game refinement theory and reinforcement schedules was conducted on 14 popular animations. It was found that different plot "placement" through the episodic development of the animations played a huge role in their entertainment, supported entirely based on viewer's data on the Bilibili website. And a good animation, aesthetics and art are equally attractive.

Keywords: animation, game refinement theory, reinforcement schedule, entertainment;

1. Introduction

Different people have different needs to watch the animation, and even more, influential directors will think about the balance of entertainment and artistry of a work. Focusing too much on a specific expression will likely cause it to become obscure and difficult to understand. Entertainment needs to be designed. Excellent animation should be from the shape design, action design, line design, plot design, and many other aspects to entertain the creative design (Jia, 2009). It makes people want to watch more and more. Video game designers used variable-ratio (VR) reinforcement schedules to keep players wanting to play the game. In the variable-ratio reinforcement schedule, the number of responses needed for a reward varies. This conditioning is the most potent partial reinforcement schedule (Skinner, 1981).

Bilibili is a Chinese video-sharing website based in Shanghai, themed around animation, comics, and games (ACG). It gathers many people who like animation. By December 31, 2020, the average monthly active users reached 202 million, with 54 million daily active users. 190 anime were online in 2017, 183 in 2018, and 81 in 2019 as of May 6. Bilibili has a scrolling commenting system nicknamed "bullet curtain" (Chinese: *danmu*; Japanese: *danmaku*). Unlike ordinary video sharing sites that only display in a dedicated comment section under the player, with this feature, viewers can post their comments while watching an anime, which will be instantly displayed with sliding over subtitles when all viewers are watching the anime; thus, increasing the interactivity among viewers. In addition to "bullet curtain," you can also donate coins and comments to your favorite anime at Bilibili. Therefore, the data of their "bullet curtain," comments, and coins of each episode on the Bilibili website (retrieved from https://www.bilibili.com/), are collected to measure the relative popularity of an anime in recent years while analyzing the impact of entertainment in animation.

2. Game Refinement Theory, Variable Ratio Schedule, and Application in Animation

Quantifying entertainment impacts in games have been previously conducted through the game refinement (GR) theory (Iida et al. 2004). According to the GR theory, it is assumed that every game's progress was encoded and transported in the minds. Although the behavior of the physics of information in the brain is unknown, the acceleration of information progress is likely subject to physics' forces and laws. The GR theory has been applied to many different games for quantifying their entertainment impacts (Iida and Khalid, 2020). From a player's perspective, the game outcome is a function of time t (i.e., number of possible moves or successful score) and the game process as solving game outcome uncertainty $x'(t)$, then (1) is obtained.

$$x'(t) = \frac{n}{t}x(t) \qquad (1)$$

The parameter n $(1 \leq n \in N)$ is the number of possible options, and $x(0) = 0$ and $x(T) = 1$. Here $x(T)$ stands for the normalized amount of solved uncertainty. Note that $0 \leq t \leq T$, $0 \leq x(t) \leq 1$. Equation (1) implies that the rate of increase in the solved information $x'(t)$ is proportional to $x'(t)$ and inverse proportional to $x(t)$ and inverse proportional to t. Then, (2) is obtained by solving (1).

$$x(t) = (\frac{t}{T})^n \qquad (2)$$

In most games, the game's total length is significantly different for players with different levels. Assuming that the solved information $x(t)$ is twice derivable at $t \in [0, T]$, then the second derivative of (2) indicates the accelerated velocity of the solved uncertainty along with the game progress. It has been found that sophisticated games have a similar GR value located at the zone of $GR \in [0.07, 0.08]$. It is expected that such sophistication also existed in the domain of animation.

It is essential to convey enough information to the viewer to catch people's attention before losing interest. Moreover, the later episodes' development and pace depend on effective information delivery speed; too fast or too slow can lead to many problems. After a continuous climactic or depressing plot, the plot's transition should give the viewer some time to rest and adjust. Continuous high emotion or depression will also make the viewer feel tired and less receptive to the information.

Engaging and attractive episodes were loved and discussed by everyone in terms of content and graphics quality. When an attractive episode was perceived, the "bullet curtain" was sent by viewers to entice interactions with other viewers. Alternatively, viewers can also leave comments to express their feelings or donate coins to show their support. With different climax, trough, and transition, the plot development curve can be qualified as an elegant plot or storyline. Hence, GR's measure in animation can be roughly formulated as (3).

$$GR = \frac{\sqrt{G}}{T} = \frac{\sqrt{Attractive\ \ episodes}}{total\ episodes} \qquad (3)$$

Adopting the principles of the VR schedules, the parameter *N* shows the average reward frequency, where *1* $< N \in$ *R*. The animation initially consisted of moments that attract viewers (i.e., settings, atmosphere, placements, to name a few) associated with high-energy and appealing after the plot's development and buildup (Xiaohan, et al., 2020). In this study, these moments were thought of as the rewards that viewers receive during the viewing process. Based on the above principles, given *N* equals the number of episodes, then the velocity *V* is formulated as (4). Such velocity measure is utilized as the total perceived entertainment of the viewers for each animation which can be roughly associated with the average "reward" felt by the viewers.

$$v = \frac{1}{N} \qquad (4)$$

3. Data Collection and Analysis

In this paper, the data of their "bullet curtain," comments, and coins of each episode of 14 popular animations on the Bilibili website were collected, where the fourth column of Table 1 gave its combined average for each episode. By adopting a boxplot, the attractive episodes can be determined by finding the episode that exceeded the third quartiles value of the combined average episode data (Figure 1). Then, GR and velocity measures was adopted to analyze further each animation's sophistication and entertainment aspects.

Table 1. Data of the 14 popular animations from Bilibili website (ordered by game refinement values)

Name (Year First Published)	Episode(s)		Average Episodes (in millions)		G	T	GR	V
	Attractive	Total	Views	Data				
The Legend of Luo Xiaohei (2011)	1, 27, 28	28	852.14	12.92	3	28	0.062	0.3
Yi Ren Zhi Xia (2016)	13 & 24	24	1251.13	30.62	2	24	0.059	0.5
Rakshasa Street (2016)	1, 24	24	1870.54	20.55	2	24	0.059	0.5
The Disastrous Life of Saiki K.2 (2016)	1, 24	24	405.63	8.38	2	24	0.059	0.5
Land of the Lustrous (2017)	1 & 12	12	526.42	16.52	2	12	0.118	0.5
Yi Ren Zhi Xia 2 (2018)	12	12	1273.50	28.60	1	12	0.083	1
Cells at Work! (2018)	1	13	1553.69	51.15	1	13	0.077	1
Violet Evergarden (2018)	1	13	885.54	49.65	1	13	0.077	1
Scissor Seven 1 (2018)	10	10	2418.80	51.56	1	10	0.100	1
Re:Zero - Starting Life in Another World (2018)	18 & 25	25	919.24	61.21	2	25	0.057	0.5
Kimetsu No Yaiba (2019)	1, 19, 26	26	2413.54	66.02	3	26	0.067	0.3
Scissor Seven 2 (2019)	10	10	2191.80	58.56	1	10	0.100	1
Heaven Official's Blessing (2020)	1	11	2352.82	93.66	1	11	0.091	1
White Cat Legend (2020)	1 & 12	12	926.00	44.02	2	12	0.118	0.5

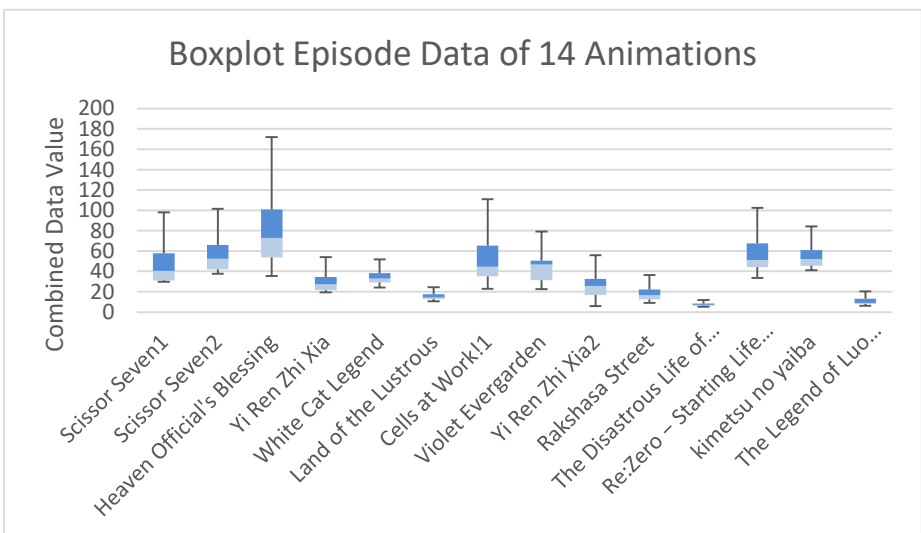


Fig. 1. The boxplot of the combined data of "bullet curtain," comments, and coins for each episodes of the considered 14 animations.

Currently, animation becomes quarterly animation or half-year production having episodes between 10~13 or 20~26. Based on the Spearman's Rank-Order Correlation (or Spearman's Rho), views very weakly correlate to GR values (ρ (12) = 0.21928, p (2-tailed) = 0.45133) and weakly correlate to V values (ρ (12) = 0.35407, p (2-tailed) = 0.21423). However, both showed positive correlations, which indicated that correlations between views and GR values (or V values) showed gradual increase trends throughout the years, implying animation production can expect more viewers with increase sophistication (GR) and entertainment (V).

The weak correlation between viewers, GR, and V value was due to several factors. Firstly, the inconsistency between appealing plot points of the animation can be observed where it was sometimes set at the beginning of the story to attract the viewer's attention (as evident from Figure 2). Meanwhile, some appealing plot was set at the end, where it was slowly built. However, such a strategy makes the whole animation the least entertaining. Such a situation can be countered by adding appealing plots in the middle and late stages to enhance the viewer's experience, as shown by Kimetsu No Yaiba and The Legend of Luo Xiaohei. This condition also aligned with the GR measure, where both were considered relatively close to the sophistication zone.

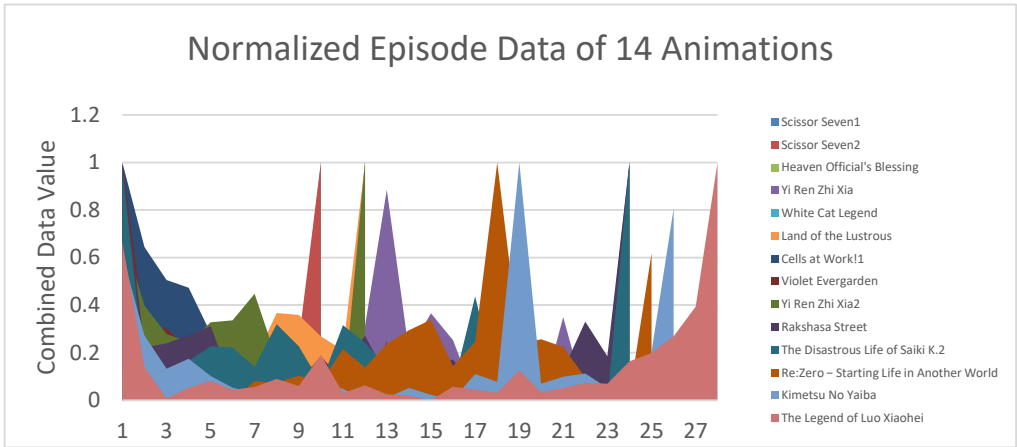


Fig. 2. The area plot of the normalized combined data of "bullet curtain," comments, and coins for all 14 animation's episodes.

4. Conclusion

Rewards and entertainment that goes hand-in-hand in the animation industry are fundamentally crucial for creative design. Making the uncertainty of rewards being expressed through media consumptions would generally be a winning idea. This study presents a preliminary study on how such reward aspect could impact the viewers' entertainment in animation retention.

It was found that each animation had different sophistication catered to different viewer's needs. Different animations had different plot placement throughout the episode's distribution which showed some correlations with the viewer's entertainment and animation production's sophistication. Hence, depending on the animation's length, appropriately placed "attractive" episodes would make the whole animation episodes entertaining. Nevertheless, further analysis of animation works provides the aesthetic and artistic attractions worth venturing in future research.

References

Iida, H., Khalid, M. N. A. (2020). Using games to study law of motions in mind. IEEE Access, 8, 138701-138709.

Iida, H., Takahara, K., Nagashima, J., Kajihara, Y., & Hashimoto, T. (2004, September). An application of game-refinement theory to Mah Jong. In International Conference on Entertainment Computing (pp. 333-338). Springer, Berlin, Heidelberg.

Skinner, B. F. (1981). Selection by consequences. Science, 213(4507), 501-504.

Xiaohan, K., Khalid, M. N. A., & Iida, H. (2020). Player Satisfaction Model and its Implication to Cultural Change. IEEE Access, 8, 184375-184382.

Zhu Jia. (2009). Qian tan ka tong dong hua de yu le she ji. art education, 11.

Analysis of Professional Basketball League via Motion in Mind

Naying Gao[a]*, Mohd Nor Akmal Khalid[b], Hiroyuki Iida[b]

[a]School of Information Science, Japan Advanced Institute of Science and Technology, 923-1211 Ishikawa, Japan
[b]Research Center for Entertainment Science, Japan Advanced Institute of Science and Technology, 923-1211 Ishikawa, Japan
*Email: gaonaying1@jaist.ac.jp

Abstract

When players play a game using strategies as players or indulge in some games as an audience, some motions such as speed, uncertainty, and unpredictability of the game during the game information process according to the different physical motions in our minds can be experienced. Such a point can be demonstrated by Basketball's popularity, where players and spectators are excited about every attempt and every shot. Attempts and shots create the information progress of the basketball game, which can be measured and visualized by adopting physics' analogy. Through the data from the overall game side, these decades to discuss how game rules changed affects game itself and dynamical data from one specific game process to explore how different motions in mind affects different level teams. This study explores how basketball rules changed in recent decades and discuss specific Basketball teams' corresponding trends, and we knew that game rules changed continuously to make GR value and AD value located in the most comfortable zone to cater to the audience and players, also we explored the comparison of motions in mind of different level teams.

Keywords: NBA, basketball, motion in mind, game refinement theory

1. Background

The National Basketball Association (NBA) is the most representative basketball league, and the NBA is the most classical and popular basketball competition. Basketball rules continually improved to help basketball be more popular and acceptable for both audiences and players(https://cdn.nba.net/nba-drupal-prod/nba-rules-changes-history.pdf). As such, it is difficult to ascertain which games and the underlying reasons for the NBA games to be popular and highly engaging to audiences.

GR theory is a measure of games' entertainment, attractiveness, and sophistication (Iida et al. 2004). Such a theory evolved based on physics' analogy by establishing velocity (v) and mass (m) as success rate and difficulty in games, respectively, via zero-sum assumption (Iida and Khalid, 2020). According to the GR theory, it is assumed that every game's progress was encoded and transported in the minds. Although the behavior of the physics of information in the brain is unknown, the acceleration of information progress is likely subject to Newton's Classical Mechanics such as physics' forces and so on. (Milton, Graeme W., and John R. Willis. (2007).

From a player's perspective, the game outcome is a function of time t (i.e., number of possible moves or successful score) and the game process as solving game outcome uncertainty $x'(t)$, then (1) is obtained. With the adoption of physics analogy, various physic-related measures can be determined (Iida and Khalid, 2020), such as velocity, momentum, potential energy, and force as given by (2) – (5). Here, G and T denotes the successful goals and total shoots, respectively. In the current context, GR (game refinement value) represents the square root of acceleration and AD (addiction value) represent the cube root of jerk which is the second and third derivatives of (1), respectively.

$$x'(t) = \frac{n}{t}x(t) \quad (1)$$

$$v = \frac{G}{T} \quad (2)$$

$$P_1 = mv \quad (3)$$

$$E_p = 2mv^2 \quad (4)$$

$$F = ma = \left(1 - \frac{G}{T}\right)\frac{2G}{T^2} \quad (5)$$

Based on previous study by Iida and Khalid (2020), it was found that GR and AD of Basketball game was 0.073 and 0.046, respectively. These values indicate that basketball can be categorized as sophisticated and highly engaging game (Naying et al. 2020). In this study, the rule change and analysis of the game process information of professional basketball games were analyzed. Different motion in mind measures between strong and weak teams were also analyzed based on such measures.

2. Motion in Mind and National Basketball Association (NBA) Game Analysis

From the previous database on the rules (https://cdn.nba.net/nba-drupal-prod/nba-rules-changes-history.pdf) and data collected these decades (https://stats.nba.com/), both GR and AD values evolution were approaching the reasonable zones (GR of [0.07, 0.08] and AD of [0.045, 0.06]) implying that the game rules becomes more reasonable and sophisticated over the years, which explains the reason for basketball popularity over the decades (Figure 1 (left)). To further analyse the game process of NBA team, the competition between the Milwaukee Bucks (Rank 2) against the New York Knicks (Rank 26) was adopted. Preliminary analysis indicates that there is a moment in the game scoring that makes one of the team "overpowering" its opponent decades (Figure 1 (right)). However, such measures may be unreliable in some cases (i.e., tied game). As such, motion in mind measures were utilized to analyse the game process.

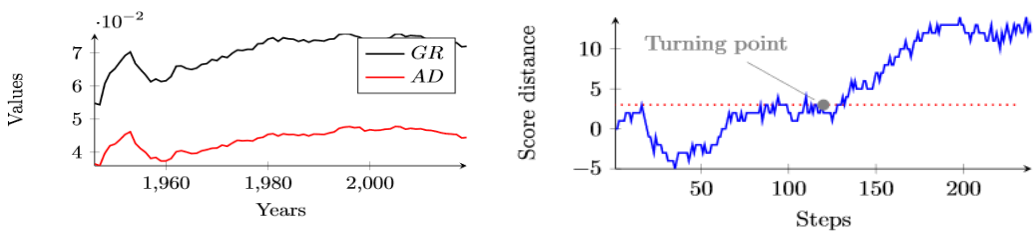


Fig.1. GR and AD trends over the recent decades of rule changes (left) and the score differences between two NBA teams in a competition that indicates there is a moment of "turning point" between win/lose the game (right)

Based on the v measures, the Milwaukee Bucks approached v = 0.5 quickly and capable enough to keep about v = 0.52. While the New York Knicks' velocity also approached v = 0.5 quickly, they cannot keep the velocity, which was overpowered in a brief time where the v = 0.43. Meanwhile, Milwaukee Bucks and New York Knicks' forces are both located at 0.004 at the end of the process, and the tendency of force is almost similar when the game ends. In physics, a force is any interaction that, when unopposed, will change the motion of an object. We used force to measure the interaction and adhesion between players with games on the psychological side, and we can see after these decades will be approached to a similar force zone. As the game progresses, our curiosity decreases, and our inertia increases; thus, our adhesion between players and the game decrease. It is reasonable to finish the basketball game when the force is located at 0.004; if the play is prolonged, the force will be too small to attract players.

However, the force of a strong team at the beginning of the game process is much higher than the weak team, which means adhesion between the game and the (strong) team is higher than the weak team during the game process. The notion of conservation of energy is proposed in this study to justify such conjectures.

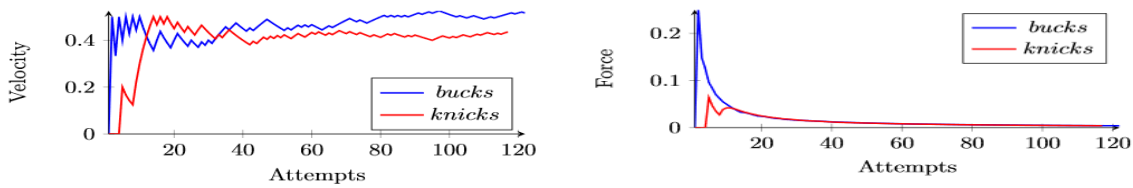


Fig. 2. The velocity (left) and force (right) of competition game between Milwaukee Bucks versus New York Knicks

3. Energy Conservation and Play Expectation Gap

In this study, objective momentum, given by (3), is regarded as the actual motion to play the game (P-1). Based on the analogy of energy conservation, the notion of subjective momentum can be derived from the expected motion of play (P2) by considering the relations of (3) and (4), then the subjective momentum can be defined as (6). Such a measure is considered to represent the gap between concentration and expectation of play experience.

$$P_2 = 2mv^2 - mv \quad (6)$$

After analysis, the momentum P_1 of Milwaukee Bucks and New York Knicks are in 1/4 at the end of the process, and the strong team approached 1/4 in a short time and kept the most reasonable momentum to concentrate on the game in a long time (Figure 3). However, the weak team's momentum needs to take longer to locate around 1/4, which means the weak team needs to use longer steps to concentrate on the game. Simultaneously, the strong team spent shorter steps to let E_p stable and located in 1/4, a strong team can reach his expectation quickly, but the weak team needs to spend a longer time. A strong team has a higher E_p value than a weak team, which means that a strong team has high expectations.

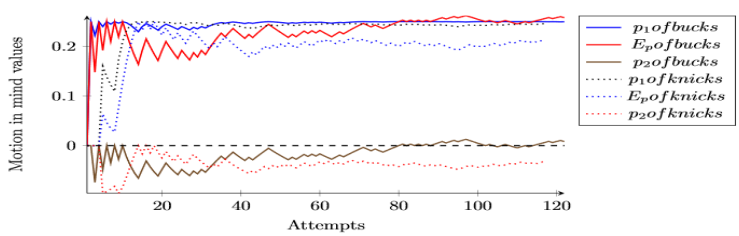


Fig 3: The objective momentum (P_1), subjective momentum (P_2), and potential energy (E_p) measures between Milwaukee Bucks versus New York Knicks

Furthermore, the P_2 of Milwaukee Bucks changes from positive (0.008), and the momentum of the New York Knicks is always negative (-0.031). The P_2 of a strong team is higher than the weak team, which means that the strong team has a high risk-taking engagement, and the process of a strong team is fascinating. The

distance between 0 and P_2 would be the distance between expectation and concentration, the momentum of the mind's motion if P_2 approached 0 means a harmonious balance between expectation and concentration, which makes the team feel comfortable. Meanwhile, the gap between P_2 and 0 of the strong team is smaller than the weak team, which means that the differences between objective and subjective momentum are minor, and the pulling force of the game and mind is small so that the player can feel more comfortable.

4. Discussion and Conclusion

This study had collected data on NBA rules change throughout the recent decade. The NBA rules have developed, updated, and explored constantly; after these years, the current rules make the GR value closer and closer to the most reasonable zone 0.07-0.08, and the AD value is getting closer and closer to the level of 0.045-0.06. With physics in mind, it can be observed that the v and m both approached 1/2 and momentum P_1 approached 1/4, the rules of the NBA keep the game is fairer and fairer after these decades. The NBA game's force is around 0.003, which gets close to a reasonable zone for basketball, making players feel exciting and attractive and attract players most reasonably. Moreover, it is the best time to finish the game; if longer, the force will be smaller and cannot attract players.

Moreover, the E_p of NBA is located around 1/4 to make the expectation of NBA more specific, and the momentum of mind's motion (P_2) stabilized around 0 after getting higher, it meant that the momentum of mind's motion is getting higher, and NBA is fascinating, and players have risk rate to take the engagement. Furthermore, this study considered the physics in the mind of different level players from the game's side; it can be observed that a strong team reached the $v = 1/2$ in a short time while re keep the velocity. However, the weak team needs to spend a longer time to reach the velocity and has a lower ability to keep the velocity, and the velocity of the weak team is a little lower than $v = 1/2$.

While both teams' forces are 0.004, the strong team has a higher force in mind during the game's initial process, which means the strong team had better adhesion. The energy conservation of different teams was proposed to analyze the momentum of the mind and game, where the strong team reached the $P_1 = 1/4$ and $E_p = 1/4$ in a short time. Thus, the strong team can reach his expectation and concentration quickly. Moreover, the velocity and momentum, the potential energy of a strong team are higher than the weak team, which means the ability and concentration, the expectation of a strong team are higher than the weak team. Meanwhile, the P_2 of a strong team is higher than the weak team, the strong team has a higher magnitude of engagement, and the process of a strong team is fascinating. Also, the distance between P_2 and 0 is associated with a sense of balance, making the strong team felt more comfortable because the distance is smaller than the weak team.

Game refinement theory and motions in mind concept can commonly be used to measure and analyze the sophistication of game from the overall game side. Also, for the dynamical game process, we can analyze the details of each team or each player, from objective sides and subjective sides, to know how instantaneous feelings during the game process based on different game results such as scores.

References

Hiroyuki Iida and Mohd Nor Akmal Khalid. "Using games to study law of motions in mind." *IEEE Access*, 8:138701–138709, 2020

Iida, H., Takahara, K., Nagashima, J., Kajihara, Y., & Hashimoto, T. (2004, September). "An application of game-refinement theory to Mah Jong." In *International Conference on Entertainment Computing* (pp. 333-338). Springer, Berlin, Heidelberg.

Milton, Graeme W., and John R. Willis. "On modifications of Newton's second law and linear continuum elastodynamics." In *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 463.2079 (2007): 855-880.

Naying, G., Primanita, A., Khalid, M. N. A., & Hiroyuki, I. I. D. A. (2020, May). “A Key Factor to Maintain Engagement: Case Study Using ‘Login System’.” In *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)* (pp. 492-497). Atlantis Press.
NBA stats per game. https://www.basketball-reference.com/leagues/NBA_stats_per_game.html. (Access on 1 March 2021).
NBA rules changes history. https://cdn.nba.net/nba-drupal-prod/nba-rules-changes-history.pdf (Access on 1 March 2021)

The Entertainment Appeal of Rhythm Games

Yuexian Gao[ac], Chang Liu[a],Mohd Nor Akmal Khalid[ab]*, Hiroyuki Iida[ab]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, 923-1211 Ishikawa, Japan
[b] Research Center for Entertainment Science, Japan Advanced Institute of Science and Technology, Nomi, 923-1211 Ishikawa, Japan
[c] Hebei University of Engineering, Handan, 056001, China
* Email: akaml@jaist.ac.jp

Abstract

Game refinement theory (GR) has been used to measure game sophistication, which could make the elements of games apart mathematically to see where the game can be improved. Over the development for almost a decade, GR theory also refined itself and extended to a broader application. Rhythm game sometimes can win the popularity of players while sometimes not. This study selects 6 rhythm games as the target. Result shows that rhythm games should be improved at clear conditions, game length, and combinations of buttons.

Keywords: Music; Rhythm games; game refinement theory

1. Introduction

With the rapid development of science and technology, people's entertainment activities are becoming more diverse. Music has also been integrated into people's lives in different forms with the change of times. A rhythm game is a music-themed action video game that challenges the player's sense of rhythm and the ability to react with your fingers or body. These games often mimic the drumbeats and melodies of dance or instrumental music. The player presses the corresponding button according to the melody stream on the screen and performs the corresponding action. The score is determined by the player's reaction and the rhythm flow of the game.

The first rhythm-based electro-mechanical arcade game was considered as it is from japan (Webster, 2008). The precursor to Dance Dance Revolution, Dance Aerobics (or Dance Studio as it was known in Japan), had players use the NES Power Pad to mimic an on-screen instructor who moved to the music. Kasco created a rhythm-based electro-mechanical arcade game in the early 1970s. The first true rhythm game is PaRappa the Rapper, credited in 1996 (Parkin, 2009). As the Guitar Hero and Rock Band get their popularity with the gamers, the rhythm game also wins its position (Matthews, 2009). Although the rhythm game has gone through a series of setbacks (Stuart, 2011), with the development of the rhythm game, people find it can help us keep healthy (Bégel, 2017). In the past two years, due to the ravages of COVID-19 globally, people must live at home and gain more consciousness of health; rhythm games such as Just Dance (2020 & 2021), which can bring fitness effects at home, have ushered in a new highlight moment.

Mathematics can be used to analyze musical rhythms, study the sound waves that produce musical notes, explain why instruments are tuned, and compose music (Garland, 1995). Game Refinement Theory was proposed to describe the information process during entertainment activities (Iida, 2004). It was widely used to quantify the sophistication of a game. Over the development for almost a decade, the mathematical measurement extended to a broader application. In this study, 6 rhythm games were selected as a research benchmark to explore musical games' motion in mind. The study would reveal part of the phenomenon for music study as well.

2. Game Refinement Theory Meets Rhythm Games

Game refinement (GR) theory (Iida, 2004) is used to measure the sophistication (i.e., the balance between luck and skill) of a game. It provides an innovative game dynamic view by evaluating the process of a game, simulating the game outcome's uncertainty as a kind of force in mind (Sutiono, 2014), which corresponds to force in nature based on Newton's second law.

From the perspective of players, information of the game outcome is a function of time t (in board games would be the number of possible moves). By considering the process of game as a process of solving game outcome uncertainty (or obtained certainty) $x'(t)$, then (1) is obtained.

$$x'(t) = \frac{n}{t}x(t) \quad (1)$$

The parameter n $(1 \leq n \in N)$ is the number of possible options, and $x(0) = 0$ and $x(T) = 1$. Here $x(T)$ stands for the normalized amount of solved uncertainty. Note that $0 \leq t \leq T$, $0 \leq x(t) \leq 1$. Eq. (1) equation implies that the rate of increase in the solved information $x'(t)$ is proportional to $x'(t)$ and inverse proportional to $x(t)$ and inverse proportional to t. Then, (2) is obtained by solving (1).

$$x(t) = (\frac{t}{T})^n \quad (2)$$

In most board and continuous movement video games, the game's total length is significantly different for players with different levels. We assume that the solved information $x(t)$ is twice derivable at $t \in [0, T]$. The second derivative here indicates the accelerated velocity of the solved uncertainty along with the game progress. After the second derivative, GR's measure can be formulated in its root square form, as given by (3). It has been found that sophisticated games have a similar GR value located at the zone of $GR \in [0.07, 0.08]$ as shown in Table 1. The game progress pattern or its difficulty is essential concerning the player's engagement.

$$GR = \frac{\sqrt{n(n-1)}}{T} \quad (3)$$

Table 1. Various popular games with measurement of game refinement theory

Game	GR
Chess	0.074
Shogi	0.078
Go	0.076
Basketball	0.073
Soccer	0.073

While we play a song of the rhythm game, we should adjust the finger for the next hit in all time. Therefore, this study considers that every time we touch the button or prepare for the next hit as the decision progress. The gameplay depends on the song, one song finished, and the player is achieving an exact condition that means a clear stage. A player should concern the hit time for every button. Hence, according to the rule, the parameter n in rhythm game progress can be given by possible options of each step, and game length is given by the average touch time of one game. Therefore, the measure of GR is calculated as Equation (4).

$$GR = \frac{\sqrt{possible\ options\ of\ each\ step * possible\ options\ other\ than\ the\ chosen\ move}}{the\ average\ touch\ time\ of\ one\ game} \quad (4)$$

This study focuses on six popular rhythm games (Table 2). The rule of the game is similar. When you play a song, music bricks fall towards each of the judging areas on the screen in time with the song's rhythm. The players should hit the point as they highlight each judging area. The condition of the button and clear condition are different among each game.

Table 2. Basic information of the target games

Games	Year	Platform	Number of buttons	Button reaction	Clear condition
Dance Dance Revolution (DDR)	1998	Arcade	4	3	Alive until the end
Beatmania IIDX (IIDX)	1999	Arcade	5	4	Alive until the end
Jubeat	2008	Arcade	8	2	Alive until the end
Love Live! School idol festival (SIF)	2013	Mobile	9	2	Alive until the end
The Idolm@Ster Cinderella Girls (CGSS)	2015	Mobile	9	3	Score > 700,000
Nostalgia	2017	Arcade	16	2	Score > 700,000

The game selected have two kinds of winning condition. First is the lifeline system. If the player misses any of the stream blocks, the lifeline decreases. The lifeline fell into 0 means losing the game. DDR, CGSS, IIDX, SIF incorporate the lifeline. Another is the scoring system, and there exists a required score for clearing the game; the game's score should be not less than the limit. The Jubeat and Nostalgia belong to this type.

The reactions of the button are also different. The DDR has three reactions: step, keep step, and avoid step; CGSS has four reactions: note, long note, flick, and slide; IIDX, SIF, and Jubeat have two reactions: touch and keep touch; And the Nostalgia has three reactions: touch, keep touch and flick.

To compare the game in the same condition. This study, based on game setting, analysis the result on the same level. According to the method, this study recorded the high-level rhythm gameplay and collected the gameplay statistic. Five people have played the game ten times for CGSS and SIF. Another data was from the recorded video of the 9th KAC (KONAMI Arcade Championship). The KAC divided the players into a different group and played two or three songs for one competition based on the final scores. This study collected 20 champion candidates of data in one rhythm game.

3. Experiment Result and Analysis

The result shows that the GR value of the rhythm game stays relatively lower with fewer buttons. It indicates that the players should pay more attention to consider the other button and the next hit when players were hitting a point when the number of buttons was increased. Hence, the game's sophistication involves some accretion levels to be exciting for the rhythm games.

Table 3. Experiment results

Games	Sample size	Total touch times	Number of buttons	Button reactions	Possible combinations	Average game length	GR value
Dance Dance Revolution (DDR)	20	16315	4	3	12	815.75	0.014
The Idolm@Ster Cinderella Girls	90	61718	5	4	20	685.7555556	0.028
Beatmania IIDX (IIDX)	20	27800	8	2	16	1390	0.011
Nostalgia	20	27361	9	2	18	1368.05	0.012
Love Live! School idol festival (SIF)	90	61518	9	3	27	683.5333333	0.038
Jubeat	20	18458	16	2	32	922.9	0.034

The dashed line separates a noticeable difference in Table 3, which was affected by the clear condition noted in Table 2. Those with the condition of "be alive till the end" have a small GR value, which means they are

much more complex than those "clear the stage with enough scores." Therefore, the upper four in the Table 3 would attract more skillful players but might lose their popularity among regular players.

Also, the target rhythm games have a lower value to meet the GR perfect zone. The reason is that rhythm game generally has a long game length, making the game much more challenging than most sophisticated game design (Iida, 2020). This condition could hint to this genre that reducing the game length would create a more pleasing experience for most players, improving these games' popularity. The fact is the latest rhythm games released, like Just Dance 2021, have shown the tendency to be much shorter. Future research should focus on the evolution of the length change in rhythm to further verify this finding.

Furthermore, possible combinations of the button number and button reactions also play a role in making the rhythm game engaging. Further study should also pay attention to a proper combination to make the game more sophisticated.

4. Conclusion

In this study, we overview the model of game refinement theory. It introduces a reliable model for evaluating the attractiveness of games and their sophistication. This study creatively applies the game refinement theory for modeling to determine the balance of attractiveness and excitement of the rhythm game. It found that conditions set for winning affect the expected experience of the player. In addition, life bar setting in the rhythm game would be too challenging for regular players.

Moreover, a shorter game length might be better to balance the game setting and attracts more players. Lastly, combinations of button numbers and button reactions should also reach a proper value to add more uncertainty and interest to the game. Adjusting the game parameters to reach the GR perfect range, as a reference for the design work, would considerably shorten and facilitate game numerical system design and game optimization. Later work should collect more data to make the study much more convincing.

References

Bégel, V., Di Loreto, I., Seilles, A., & Dalla Bella, S. (2017). Music games: potential application and considerations for rhythmic training. Frontiers in human neuroscience, 11, 273.

Garland, T. H., & Kahn, C. V. (1995). Math and Music: Harmonious Connections. Dale Seymour Publications, PO Box 10888, Palo Alto, CA 94303-1879.

Iida, H., Khalid, M. N. A. (2020). Using games to study law of motions in mind. IEEE Access, 8, 138701-138709

Iida, H., Takahara, K., Nagashima, J., Kajihara, Y., & Hashimoto, T. (2004, September). An application of game-refinement theory to Mah Jong. In International Conference on Entertainment Computing (pp. 333-338). Springer, Berlin, Heidelberg.

Matthews, Matt (2009). Analysis: Guitar Hero Vs. Rock Band – Behind The Numbers. Gamasutra. https://www.gamasutra.com/php-bin/news_index.php?story=25739.

Parkin, Simon(2009). Rhythm Paradise Review. EuroGamer. https://www.eurogamer.net/articles/rhythm-paradise-review

Stuart, Keith (2011). Guitar Hero axed: five reasons why music games are dying. https://www.theguardian.com/technology/gamesblog/2011/feb/10/guitar-hero-axed.

Sutiono, A. P., Purwarianti, A., Iida, H. (2014, July). A mathematical model of game refinement. In International Conference on Intelligent Technologies for Interactive Entertainment (pp. 148-151). Springer, Cham.

Webster, Andrew (2008). "Roots of rhythm: a brief history of the music game genre." Retrieved from https://arstechnica.com/gaming/2009/03/ne-music-game-feature.

Digital Game Design for Physics Education

Nor Aidatul Ismail[a], Azrulhizam Shapii[b*]

[a,b] Faculty of Information Science & Technology, UKM Bangi, 43600, Malaysia
* Email: azrulhizam@ukm.edu.my

Abstract

Malaysia's Ministry of Education confirmed that the number of students who enrolled in school's science stream option declined in the year 2018 than years before that. Physics is listed as one of the must-enroll subjects under STEM education. There is some research done previously stating that students had some negative perceptions towards Physics. Therefore, teachers or educators have to develop strategies to trigger their students' interest and motivation toward this subject. Among the strategy is the game-based learning approach. However, there are some challenges that the teachers have to encounter even though they had some interest in the approaches, mainly due to the lack of basic programming skills. This research was initiated to assist teachers in exploring the concept of game development without any programming background. This research is also aimed to highlight some free and open-source game engines available in the market. Using the Game Development Life Cycle approaches, this research explored three easy and free software to develop physics education games. The research aims to help teachers utilize the free and easy software and build their games without any basic programming skills.

Keywords: Game-based learning; digital game, teaching, and learning; STEM; Physics education

1. Introduction

Physics is one of the subjects listed under STEM education. Based on research conducted by the Faculty of Science, Universiti Putra Malaysia, students enrolled in the science stream had some negative impressions on Physics. Ministry of Education (MOE) also confirmed that the number of students enrolled in the STEM stream has been inclining by years; based on facts, 49 percent of students enrolled in 2012, but the number reduced to 44 percent in 2018. Therefore, various programs and events have been initiated to spark students' interest in understanding STEM subjects. Mazlin and Iksan (2018); Salleh and Halim (2016), based on the research on students' motivation against Physics subjects, stated that students encountered problems understanding the concept and elements because they assumed it is too abstract and concrete. Students who consider Physics as a student are irrelevant to be learned (Saleh 2014). This kind of perception gave learned on the student's achievements, which leads to situations where they did not understand what has been taught and continuously depended on teachers without any effort to master the knowledge by themselves (Jufrida et al. 2019). We have seen teachers had to venture into technology for TnL due to the closure of schools nowadays. Lay and Osman (2018) brought a new approach to the TnL of Chemistry by introducing a digital game, KimDG, to help students understand the topic of salt. In the same study, the researchers found that the game allows students to work together to understand and create more effective, understandable ideas and concepts. Anderson and Barnett's (2013) study on America's high school students showed a positive result on the Physics digital game approach, Supercharged! Rather than conventional TnL. The game-based learning approach is more likely to impact students' understanding and mastery of an impact positively. The game-based learning approach is more relevant, either digitally or not, compared to conventional learning (Hafis & Supianto 2018). However, the digital game-based learning approach could be challenging, especially when Malaysia's teachers mostly came from a non-technical background (Mat Diah & Yahya 2009).

The developers usually came from a technical background with basic coding and programming knowledge to develop a digital game. However, multiple game engines are available in the market for people to utilize, however, with some programming language mastery requirements (Chamillard 2007) such as C++, C#, or Java

programming. However, this kind of requirement makes digital games much harder for some people, especially teachers. This is due to their non-technical background and making it more complex for teachers to develop their games for the subjects. According to Mohd Hashim and Mat Diah (2016), many people have an interest and want to develop their games. However, it seems impossible for them due to no skills in programming. Even someone with programming experience also finds it hard to understand and master the skills (Mohd Hashim & Mat Diah 2016). Therefore, any game engines available without any programming skills are most likely to be utilized by people with no programming background. This study has been done to highlight the concept of non-programming game development for teachers to develop their games. We have determined multiple free and non-programming software or game engines that teachers can explore within the study. The study also did develop some simple game prototypes using the selected game engines and tested the prototype on some secondary school students that enrolled in Physics subjects around Malaysia.

2. Research Background

Physics, in general, is a subject that requires students to search for the answer for the issue of "why" and "how" certain phenomena happens around people (Saleh 2014). However, according to Saleh (2014), some students characterized Physics as a complicated subject due to the complexity of learning, requiring students to understand various formulas, calculations, and concepts. Meanwhile, in their study, Abdul Kadir et al. (2016) stated that some students had a perception that Physics is only for the intelligent student who makes Physics labeled as an "elite" subject in school. The negative perception indirectly affects students' motivation and disrupts their performances. Another study by Meng et al. (2014) mentioned that some students consider physics irrelevant. In summary, students' attitudes and acceptance towards certain subjects sometimes depend on their thought, which leads to some early perception without experiencing it first.

Siong and Osman (2018) mentioned in their study, the application of serious games in TnL could increase the students' collaboration and problem-solving skills (Antunes et al. 2012) while interacting with the game. Game-based learning is increasingly giving good impact when being used in TnL. Some examples of digital games that were successfully implemented in classroom TnL are Sim City and Civilization III. According to Squire and Jenkins (2003), students can learn the economy, social hierarchy, and politics in the game's population. Civilization III allows students to explore things that can help in the civilization played, such as searching for life supplies, generating economy, and the civilization's growth within the gameplay (Brom et al. 2010; Squire & Jenkins 2003).

To produce an exciting game that can be integrated into TnL, the design and development should be given more attention. This is for the game developed to generate interest and be a tool to deliver the knowledge effectively and innovatively. Several elements can potentially be a guide to build a digital game for education. According to Novak (2011), the first element is genre. The genre chosen by developers should be according to their targeted market, and it should be determined so that the following development process shall be continued. The second element to be considered is the visual and audio design (Plass et al., 2015). Without interesting visual and pleasant audio background, it could be hard to maintain its interest in the game. According to Plass et al.(2015), the visual design includes the characters, information, animations, etc., portraying the gameplay itself. At the same time, the third element is the gameplay and motivational value (Buckley & Doyle 2016; Plass et al. 2015). Motivational elements are essential in order to guide the player to find the purpose of the game itself. The motivational element includes players' rewards upon completing the mission (Plass et al., 2015). Lastly, the most crucial element is targeted learning translated into digital games for education (Plass et al., 2015). This element has to be the foundation for every educational game development, not to be diverted from its original goal: to educate. In their study, Liu and Chen (2013) mentioned that games make the learning session more active and positively impact the students by targeting the fun side and its value.

3. Method

This study adopted qualitative research while using the Iteration Game Development Life Cycle approach in game development. There are four phases altogether. This approach is chosen as it is suitable for the development cycle of the educational game. It is because the cycle could be revisited anytime, according to the syllabus. The details on the research development model are as per Figure 1.

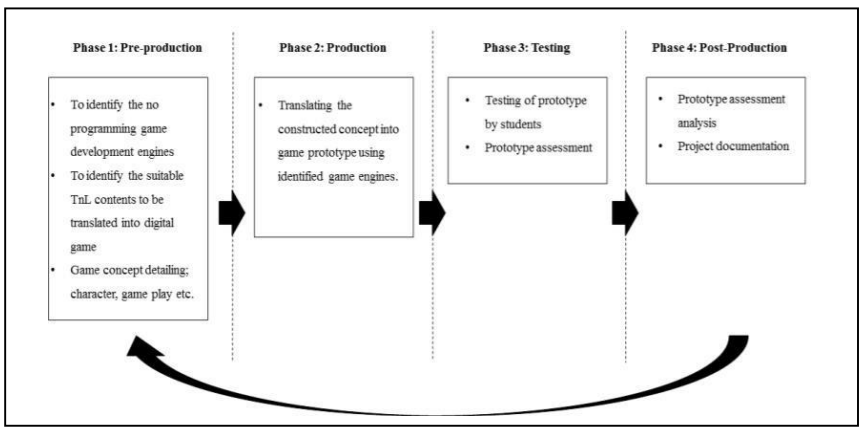


Fig. 1. Methodology

4. Design and Development

Construct 2 has been used to develop a puzzle platformer 2D game. The game is named Physics Ninja!. Physics Ninja was built based on visual programming, which is the "drag and drop" concept. The game focuses on one character: a ninja who had a mission to find the mission character, collect keys, gems, and most importantly, answer all quizzes given by the mission character. Gems and keys will be the determining marks for each student upon completion. The main character and enemy encounter will deduct the count of gems while two gems acquired for each enemy were destroyed. The process of visual programming and example of gameplay interface can be seen in Figure 2 below.

5. Conclusion

This study was designed to identify either game-based learning can potentially be implemented in Physics education with the development that requires no programming. This study focuses on assisting and guiding teachers to explore various technology tools, especially using the game-based learning approach. In this study, we have identified how digital games can be implemented, an alternative tool for student assessment. Next, in this study also we have identified multiple no programming game engines available, free for teachers to utilize and build their own game using the concept of visual programming. Lastly, in this study, we managed to identify students' assessment of the game and their acceptance of the game-based learning approach in Physics education through the development of the prototype. In conclusion, the study hopes to bring new opportunities for teachers to provide various teaching tools for Physics education and increase students' interest in the subject.

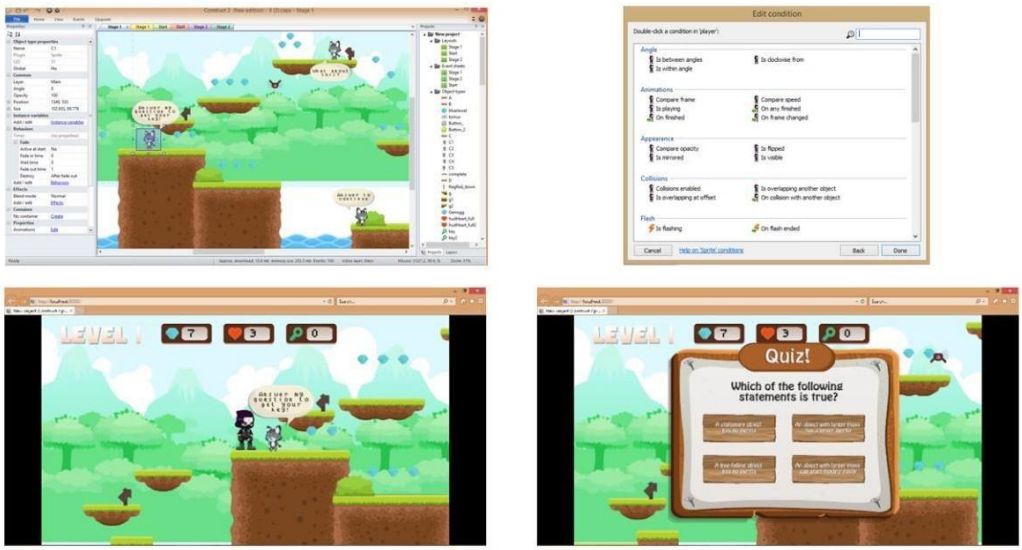


Fig. 2. Process and game interface example

References

Abdul Kadir, M.N.B., Abdul Karim, M.M. & Rahman, N.A. 2016. Sikap Pelajar Terhadap Pembelajaran Fizik dan Hubungannya Dengan Pencapaian Dalam Kalangan Pelajar Sains. *Jurnal Personalia Pelajar 19*: 31–51.

Ahmed Qasem, A. & Viswanathappa, G. 2016. The teachers' perception towards ICT integration: Professional development through Blended learning. *Journal of Information Technology Education: Research 15*: 561–575. https://www.informingscience.org/Publications/3562.

Anderson, J.L. & Barnett, M. 2013. Learning Physics with Digital Game Simulations in Middle School Science. *Journal of Science Education and Technology 22*(6): 914–926.

Brom, C., Šisler, V. & Slavík, R. 2010. Implementing digital game-based learning in schools:augmented learning environment of 'Europe 2045.' *Multimedia Systems 16*(1): 23–41. http://link.springer.com/10.1007/s00530-009-0174-0.

Hafis, M. & Supianto, A.A. 2018. Mobile game design for learning chemical bonds with the endless run approach. *International Journal of Interactive Mobile Technologies 12*(8): 104–112.

Jufrida, J., Kurniawan, W., Astalini, A., Darmaji, D., Kurniawan, D.A. & Maya, W.A. 2019. Students' attitude and motivation in mathematical physics. *International Journal of Evaluation and Research in Education 8*(3): 401–408.

Multi-points Navigation for Autonomous Robot in Duct Environment

Ghassan Jasim AL-Anizy[a], Khairunnisa' Ahmad Shahrim[a], Mehak Raibail[a], Abdul Hadi Abd Rahman[b]*

[a]Machine Learning and Vision Lab, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
[b]Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia.
* Email: abdulhadi@ukm.edu.my

Abstract

Navigation in a duct environment is challenging due to several factors such as size, shape, and lighting. Various research has been initiated to discover the best approach to allow autonomous robot navigation using appropriate path planning algorithm. However, most papers focused on the sewer environment, which is different compared to the air duct environment. While the lighting is considered ineffective, robot navigation should rely on a laser range finder and multi-point references. This paper we study the effectiveness of rapidly exploring random tree (RRT) and path planning algorithms for a mobile robot in a simulated duct environment. The RRT algorithm is chosen due to its biased towards unexplored regions. The experiment is conducted using a single robot in Robot Operating System (ROS) framework to evaluate the capability of multi-point robot navigation. It is observed that successful navigations are achieved in repeated experiments. Future work will concentrate on multi-agent system and optimizing different path planning algorithm.

Keywords: RRT; Path Planning; Duct Environment; Autonomous Robot

1. Introduction

Navigating mobile robots in a duct environment is challenging due to several factors, such as the small size of ducts (Koledoye et al. 2017), various shapes and dimensions (Chataigner et al. 2020). The ability to overcome all these constraints allows autonomous mobile robots to perform various tasks such as surveillance or cleaning tasks. Exploration techniques based on frontiers (Wang et al. 2011) are widely used for robotic exploration to direct robots to frontier edges. However, there are no general techniques that can ensure successful autonomous navigation in all kinds of environments, especially in the duct.

Exploration of duct environment for mapping requires high computational due to the existence of various junctions and long tunnels. Senarathne et al. (2013) highlight the benefit of occupancy grid map used to store the map as the map grows larger, processing it can consume more and more computational resources. Various exploration strategy has been investigated by previous researchers using single and multiple robots. Other widely used navigation components, such as the simultaneous localization and mapping (SLAM) module (Grisetti et al. 2007), and the route planner module, are used to evaluate the exploration strategy.

Varnell et al. (2016) suggested RRT-based exploration approaches for quick path planning and exploration in higher dimensional spaces for computing invariant sets of dimensional systems. Although various path planning algorithms are available, RRT proves that it can provide an effective solution for solving path planning in various environments. This paper investigates the effectiveness of RRT and path planning algorithm for multi-point navigation of a mobile robot in a simulated duct environment.

2. Methodology

2.1 Duct Environment

A real air duct simulation is used as a reference to develop the duct simulation environment. We develop an area of 8 meters (width) x 9.5 meters (length). The hallway size varies from 0.5 to 1-meter consist of the main hallway, connection, and blower channel. There are four junctions involves in this simulation setup. The starting point was set in the third row of the channel. Six reference points are defined in x, y, z coordinates as [-3.9,5.8,0,1.1,5.8,0,-3.9,3.2,0,1.2,3.2,0,-4,-1.7,0,1.1,-1.7,0].

2.2 Hardware requirement

Turtlebot3 Waffle is used as the main robot platform to navigate inside this tunnel due to the suitability of its height (15cm) and width (30cm). The localization of the robot utilizes the laser range finder, which ranges up to a 5-meters radius. Turtlebot3 Waffle is embedded with 2 wheels motor that helps to balance its movement. Turtlebot3 Waffle camera is used for visualization during navigation.

2.3 Mapping

This project implements a known environment where the mapping of the environment is conducted before the path planning algorithm is initiated. The mapping process is performed using the teleoperation function. The map is further improve using the Image Editor tool to clean up the unnecessary and unconnected lines. This map has been tested during the 2D navigation process to ensure the robot can move in all duct areas.

2.4 RRT Algorithm

RRT is a path planning algorithm (Zhang et .al 2018) that samples space using randomly generated points. Random points are used to extend edges in a tree-like structure, which consists of nodes and edges. One possible mode of RRT-based exploration is to make robots follow the above-mentioned tree structure as the tree structure grows. In this project, the RRT algorithm is used to explore the path planning to navigate to all 6 pre-set reference points. The RRT algorithm is chosen because it is biased towards unexplored regions, encouraging the tree to detect frontier points.

In RRT, the exploration mechanism is achieved in three modules: the frontier detector, filter, and robot task allocator, as shown in Fig. 1. The frontier detector is responsible for detecting frontier points and passing them to the filter module. The filter module clusters the frontier points and stores them. The filter module also deletes invalid and old frontier points. The task allocator module receives the clustered frontier points from the filter module and assigns them to a robot for exploration.

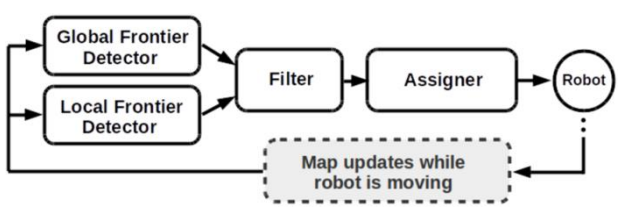


Fig. 1. RRT structure

3. Experimental Results

3.1 Simulation duct environment

The visualization of sensory data, robot position, and the duct environment using Rviz tools is showing in Fig. 2. The robot's odometry is contained linear, and orientation is observed using a terminal by echoing */odom* topic. A total of 6 minutes and 16 seconds of simulation time is obtained to complete the whole navigation to all 6 points. Our observation shows the robot required about 20 seconds to initialize its position onto the map, although the predefined origin has been defined.

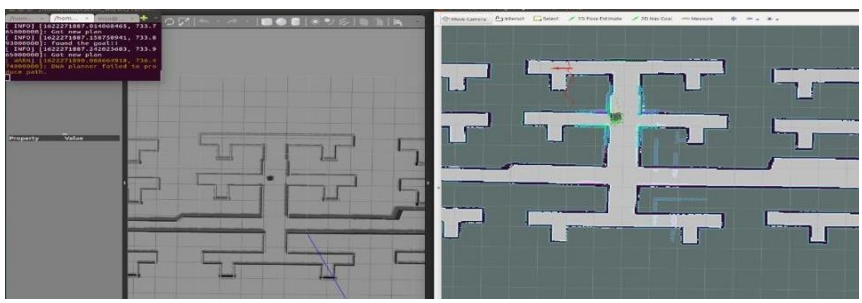


Fig. 2. Navigation to multi-goals using RRT

3.2 Path planning

Fig. 3 shows the robot navigational position from origin to all 6 references in sequence number. It is observed that the trajectory movement of the robot is not too smooth using an existing algorithm. However, all points are achieved with distance coverage of 51.6 meters and duration as presented in Table 1. It is observed that simple path planning achieves every goal faster than RRT path planning. This is because RRT path planning navigation uses a 2D Navigation arrow in SLAM for every goal, while simple path planning uses autonomous navigation. Most existing algorithm are used using Turtlebot Indigo which is suitable only in Ubuntu 14.04. Thus, we make some changes that suit Turtlebot3 in ROS Kinetic. The experiment was conducted using Ubuntu 16.04 and Turtlebot3 in ROS Kinetic. Gazebo and Rviz are the platforms that are used for simulation of the two different path planers to see their performance. To compare the performance of the two path planning strategies, we measure the duration of every goal.

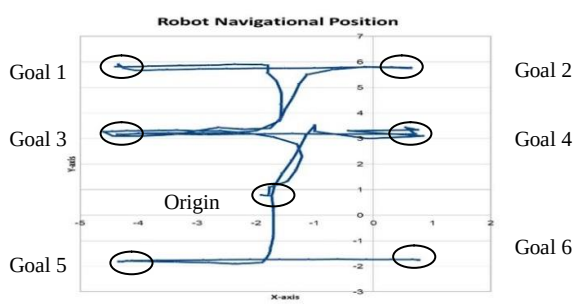


Fig. 3. Navigation of duct robotic based on six reference points.

Table 1. Distance coverage and duration from point to point

	Distance (m)	Time start		Time finish		Duration	
		Move-base	RRT	Move-base	RRT	Move-base	RRT
Origin to Point 1	13.3	1 m 44 sec	0.00	2 m 50 sec	56 sec	1 m 6 sec	56 sec
Point 1 to Point 2	4.5	2 m 51 sec	56 sec	3 m 40 sec	1 min 27 sec	49 sec	30 sec
Point 2 to Point 3	7.9	3 m 41 sec	1 min 28 sec	4 m 20 sec	4 min 22 sec	39 sec	2 min 54 sec
Point 3 to Point 4	5.0	4 m 21 sec	4 min 23 sec	5 m	6 min 57 sec	39 sec	2 min 34 sec
Point 4 to Point 5	10.3	5 m 01 sec	6 min 58 sec	5 m 42 sec	8 min 6 sec	41 sec	2 min 8 sec
Point 5 to Point 6	7.2	5 m 43 sec	8 min 7 sec	6 min 16 sec	8 min 30 sec	33 sec	23 ec

4. Conclusion

This project highlights the capability of a path planning algorithm to navigate in a duct environment successfully. The proposed algorithm is evaluated based on a duct environment using the Rviz simulator. Further improvement on the distance and duration using other algorithms is planned.

Acknowledgments

The authors want to acknowledge the Ministry of Higher Education Malaysia and University Kebangsaan Malaysia for funding and supporting this project using grant code FRGS/1/2020/ICT02/UKM/02/7.

References

Bircher, A., Kamel, M., Alexis, K., Oleynikova, H., & Siegwart, R. (2016, May). Receding horizon" nextbest-view" planner for 3d exploration. In 2016 IEEE international conference on robotics and automation (ICRA) (pp. 1462-1468). IEEE.

Chataigner F. et al., "ARSI: An Aerial Robot for Sewer Inspection," Springer Tracts in Advanced Robotics, vol. 132, pp. 249–274, 2020, doi: 10.1007/978-3-030-22327-4_12.

Grisetti, G., Stachniss, C., & Burgard, W. (2007). Improved techniques for grid mapping with raoblackwellized particle filters. IEEE transactions on Robotics, 23(1), 34-46.

Zhang, H., Wang, Y., Zheng, J., & Yu, J. (2018). Path planning of industrial robot based on improved RRT algorithm in complex environments. IEEE Access, 6, 53296-53306.

Koledoye M. A., D. De Martini, M. Carvani, and T. Facchinetti, "Design of a mobile robot for air ducts exploration," Robotics, vol. 6, no. 4, p. 26, 2017, doi: 10.3390/robotics6040026.

Senarathne, P. G. C. N., Wang, D., Wang, Z., & Chen, Q. (2013, May). Efficient frontier detection and management for robot exploration. In 2013 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (pp. 114-119). IEEE.

Varnell, P., Mukhopadhyay, S., & Zhang, F. (2016, December). Discretized boundary methods for computing smallest forward invariant sets. In 2016 IEEE 55th Conference on Decision and Control (CDC) (pp. 65186524). IEEE.

Wang, Y., Liang, A., & Guan, H. (2011, April). Frontier-based multi-robot map exploration using particle swarm optimization. In 2011 IEEE symposium on Swarm intelligence (pp. 1-6). IEEE.

contact us

M-CAIT2021 SECRETARIAT
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600
Bangi, Selangor, MALAYSIA