Data Science Research Symposium 2018

Data Analytics and its Applications

July 12, 2018

organized by

Department of Information Systems
Faculty of Computer Science & Information Technology,
University of Malaya, Malaysia

Proceedings of
Data Science Research Symposium 2018

Editors:

Dr. Vimala Balakrishnan
Wandeep Kaur
Assoc. Prof. Dr. Maizatul Akmar Ismail
Khalid Haruna

 ISBN 978-983-43491-7-2 (EPUB)

eISBN 978-983-43491-7-2

9 789834 349172

Printed in Malaysia.

This proceeding is also published in electronic format
https://umconference.um.edu.my/DSRS2018

DSRS2018 WORKING COMMITTEE

ADVISOR
Dr. Sri Devi Ravana (Head of Department of Information Systems)

CHAIRPERSON
Dr. Vimala Balakrishnan (Coordinator Master in Data Science)

SECRETARY
Assoc. Prof. Dr. Maizatul Akmar Ismail (Deputy Dean of Undergraduate)

TREASURER
Dr. Mohd Khalit Bin Othman

WORKING COMMITTEE
Wandeep Kaur
Khalid Haruna
Shahzaib Khan

LOGISTICS
Puan Rohayu Mohd Nor

TECHNICAL TEAM
Mr. Huswadi Hussain
Mr. Mohd Jalaluddin Ahmad

DEPARTMENT OF INFORMATION SYSTEMS COMMITTEE
Associate Prof. Dr. Maizatul Akmar Ismail
Associate Prof. Dr. Salimah Binti Mokhtar
Associate Prof. Dr. Teh Ying Wah
Dr Vimala Balakrishnan
Dr. Azah Anir Binti Norman
Dr Fariza Hanum Binti MD Nasaruddin
Dr Suraya Binti Hamid
Dr. Hoo Wai Lam
Dr. Kasturi Dewi Varathan
Dr. Mohd Khalit Bin Othman
Dr. Nor Liyana Binti Mohd Shuib
Dr. Norizan Binti Mohd Yasin
Dr. Norjihan Binti Abdul Gani
Dr. Sri Devi Ravana

TABLE OF CONTENT

PREFACE

The Symposium Proceedings volume contains the collection of articles of all the contributions presented during the Data Science Research Symposium, held on July 12, 2018 at the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia.

The Symposium provided a platform for the post-graduate students, academics, researchers and industry players to share their knowledge in the forms of research works and opinions in various areas of Data Science. The DSRS 2018 served as a good setting for the scientific community where 38 participants met to share and exchange ideas. The participants were mainly from the local universities including University of Malaya, Taylors University, Multimedia University, with 6 from the local industries such as Telekom Nasional Berhad (TNB), MIMOS Bhd. and The Center for Data Analytics (CADS), etc. The Symposium also attracted foreign submissions from United Kingdom and United States of America.

We would like to thank all the participants for their contributions to the Symposium program, and to this proceeding as well which contain all the 34 abstracts. We also express gratitude to the working committee, namely Ms Wandeep Kaur for monitoring participants' registrations and payments, Mr Shahzaib Khan for designing the Symposium website and Mr Khalid Haruna for compiling this proceedings volume. Finally, our special thanks to our colleagues from the Department of Information Systems for their unfailing support towards DSRS 2018.

Dr Vimala Balakrishnan
AP Dr Maizatul Akmar Ismail

Mitigation of scattered mobility of sensor's data for effective indexing in Wireless Sensor Networks

Hazem Jihad Badarneh[1], Sri Devi Ravana[1], and Ali Mohammed Mansoor[2]

[1] *Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia*
[2]*Department of Software Engineering, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia*

Corresponding Emails: hazem_jihad@siswa.um.edu.my, sdevi@um.edu.my

INTRODUCTION

Wireless Sensor Networks (WSNs) have been rapidly growing due to small nodes capabilities of sensing, mobility and computation. Moreover, the sensors are required to report their location and time frequently during transmission. In consequence, a huge data generated, which required being efficiently allocated and effectively processed. However, gathering data from various sources with diverse information and perform an effective indexing mechanism are remain challenges in terms of time and overhead. The significance of this work is to reduce the challenges that caused from transmitted data in random mobile WSN, to improve the efficiency of indexing those data.

PROBLEM STATEMENT

This study intends to solve the problem of indexing data in WSN that contains random mobile sensors. This problem results from the scattered sensor's data. Data that belong to the same source are scattered throughout the different destination, as shown in Figure1. In figure1, part A consists of a set of mobile sensors represented as S0, S1, and a set of gateways, gw1, gw2, and gw3. Each sensor transfers data to the gateway with different time period, Time A, Time B, and Time C. Inside each gateway, part B we can see data that belong to the same source scattered and not arranged inside each gateway. This problem increases the index overhead because the time for searching for data is increased in which affect directly on indexing building time. Otherwise, it increases the number of index updating operations, in case of coming data belong to the same source that needs to insert to the existing node in the index.

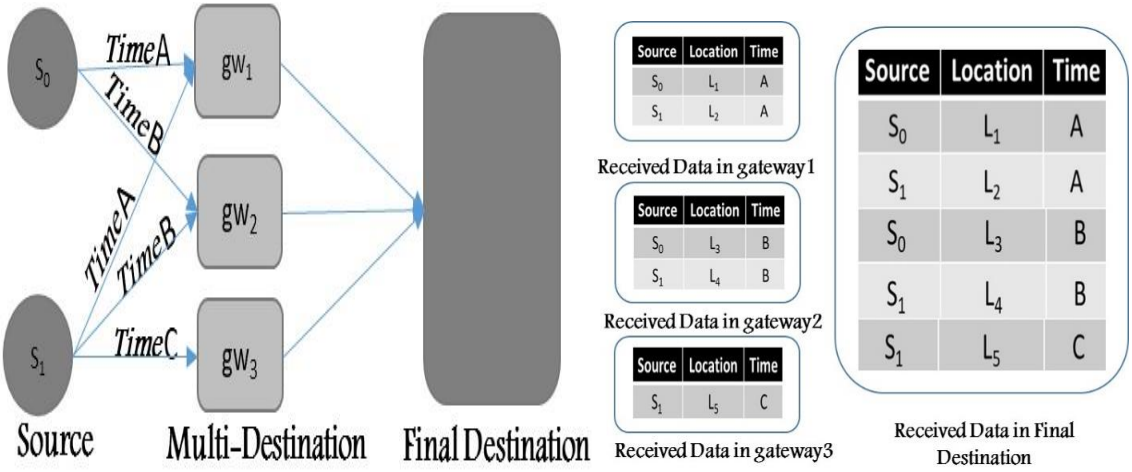


Figure 1: The problem of scattered data in the gateway

RELATED WORK

The most famous indexing techniques dedicated for mobile sensors, are R-tree and B-tree indexing techniques. R-tree (Guttman, 1984) is effective for multi-dimensions indexing, which is extended from B-tree. It divided into two parts, non-leaf node and leaf node. The main challenge in R-tree is node overflow and underflow. Node overflow is caused because of insertion, which means inserting more than a maximum number of specifying entries for indexing, the worst case of node overflow which forces the tree height to increase. Node underflow is caused by deletion operations that cause to reinsert a number of objects. The main disadvantage of R-tree has inefficiently updated operations. B-tree index (Bayer & McCreight, 2002) complies with the review of large multidimensional data related to a set of rules and operators, which provide specialized plans for search-related data. It is suitable for dealing with records of different lengths that are commonly observed in large data. However, B-tree faces high query cost because of the complexity of its structure and wastes the computing resources especially in online data indexing.

METHODOLOGY

In this work, an effective indexing method is proposed to mitigate the scattered mobility of sensor's data and to efficiently index the transmitted data with lower index overhead with respect to index building time and space cost overhead. We proposed Coalition-Energy-efficient Distributed Receiver tree (CoEd-tree) which is a two-layer indexing method based on integrating enhanced Energy-efficient Distributed Receiver (EEDR) routing protocol with chosen subsets called coalition to minimize the index building time and updating operations. It is based on initiating index step by step for all the transmitted data by using data preparation algorithm, that prepare the data before it accumulated in the main server.

FINDINGS

Experimental results show that the performance of CoEd-tree is superior to the best-known competitors R-tree and Decomposition-tree (D-tree). The performance is tested based on the following metrics: Index building time, space cost evaluation and Accuracy, with compared with the mentioned baselines

REFERENCES

Bayer, R., & McCreight, E. (2002). Organization and maintenance of large ordered indexes *Software pioneers* (pp. 245-262): Springer.
Guttman, A. (1984). *R-trees: A dynamic index structure for spatial searching* (Vol. 14): ACM.

Building Multilingual Sentiment Lexicons Based on Unlabelled Corpus

Mohammed Kaity and Vimala Balakrishnan

Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur Malaysia

Corresponding Emails: moh.kaity@gmail.com, vimala.balakrishnan@um.edu.my

Keywords*—Sentiment analysis, Text analysis, Natural language processing, Sentiment lexicon.*

INTRODUCTION

Many methods have been developed which are currently being used to implement sentiment analysis. One of the most commonly-used method for classification purposes is the sentiment lexicon. A Sentiment lexicon is specified as a list of feeling words and phrases with their sentiment classes or semantic orientations (Bravo-Marquez, Frank, & Pfahringer, 2016; Wu, Huang, Song, & Liu, 2016). In the absence of adequate training data set, the lexicon-based approach is proven more appropriate than the machine learning approach (Deng, Sinha, & Zhao, 2017). Moreover, in short texts such as social media texts, sentiment lexicons are believed to work well (Bermingham & Smeaton, 2010). They are further suitable for real-time opinion classification because their computation requirements are relatively low, (Deng et al., 2017). Furthermore, these sentiment lexicons can be employed for unsupervised (Bravo-Marquez et al., 2016; Deng et al., 2017) and supervised classification (Kiritchenko, Zhu, & Mohammad, 2014), for a given text.

Sentiment lexicons are essentially obtainable for the English language, while in numerous other languages these resources are either reduced or not available. Thus, many previous works have used translating systems to translate English lexicons to target particular languages to establish non-English sentiment lexicons (Abdaoui, Azé, Bringay, & Poncelet, 2016; Steinberger et al., 2012). In order to overcome this problem, an automatic language-independent method for producing non-English sentiment lexicons is proposed in this work. The proposed method uses existing English lexicons with unlabelled target language corpus to recognize the sentiment of the given document or word. The main contribution of this paper is a framework incorporating two available resources (seed lexicon and unlabelled corpus) to build and adapt sentiment lexicons for non-English languages.

METHODOLOGY

A corpus-based method is proposed to discover new polar words based on the following two resources: a target language corpus, and English seed sentiment lexicons. In the proposed method, the seed sentiment lexicon is utilized to specify new sentiment words in the target language corpus depending on the presence of the words next to the seed words. The method consists of four phases. First of all, seed lexicons preparation, where English sentiment lexicons are translated to the target language using machine-translating tools. Then, splitting off prefixes and affixes of the words (i.e. lemmatization). Second, the corpus is collected, pre-processed and cleaned of the links and stop words. Third, candidate words are collected and extracted from the corpus. Part-of-speech tags (POS) are then added to each candidate word in the list. Finally, the sentiment orientation of the candidate words is identified using the seed lexicon and pre-processed corpus, which will require determining the relationship between the previously known polarity words (seeds) and the "new" words (candidates). This process starts by selecting a new candidate word from the candidate word list. Then, searching the corpus for any documents that contain the candidate word. After which the seed lexicon is used to specify the polar words in those documents. The more the word is repeated in multiple documents, the more likely the word is a polarity word.

RESULTS AND DISCUSSION

Table 1 shows the performance results of the new lexicon compared to the other sentiment lexicons. The results show that the new sentiment lexicon outperformed the other sentiment lexicons, achieving 0.74 F-Measure as compared to the rest of the lexicons. The nearest lexicon with regards to the F-Measure was the Hybrid 4 lexicon which attained F-measure of 0.69. The Hybrid 4 lexicon included the new lexicon and the three seed lexicons Tra_MPQA, Tra_OL, and Tra_AFINN. However, the translated lexicons did not achieve F-measure exceeding 0.67 including NRC, which was the sentiment lexicon translated by its producers.

TABLE 1: The performance results of the UnCBSL lexicon compared with some sentiment lexicons

Lexicon	Description	Accuracy	Precision	Recall	F-Measure
Tra_OL	The translated copy of the Bing Liu's opinion lexicon	0.49	0.62	0.67	0.62
Tra_AFINN	The translated copy of the AFINN	0.50	0.68	0.73	0.65
Tra_MPQA	The translated copy of the MPQA	0.47	0.63	0.68	0.60
Hybrid 3	The combination of the three translated lexicons	0.58	0.68	0.72	0.67
UnCBSL	*Unlabelled Corpus-Based Sentiment Lexicon*, the proposed sentiment lexicon developed by our method	**0.78**	**0.78**	0.72	**0.74**
Hybrid 4	The combination of (Hybrid 3) and (UnCBSL)	0.74	0.71	0.68	0.69
Arabic NRC	NRC Emotion Lexicon	0.45	0.62	0.65	0.55
AraSenTi - Arabic	The large-scale Arabic sentiment lexicon	0.59	0.61	0.63	0.57

CONCLUSIONS

An automatic method for building non-English sentiment lexicons was proposed in this study to address many of the limitations and challenges, using two types of available and relatively cheap resources to obtain the target language Unlabelled corpus, and the English seed sentiment lexicons. The proposed method was applied to the Arabic language data where sizeable Arabic corpus was collected from the Facebook social networking site to apply the proposed method and to extract new sentiment words. The proposed method was also tested with unlabelled Arabic corpus, and three English sentiment lexicons were used as seed lexicons. The evaluation results revealed that the new Arabic lexicon improved opinion mining task, giving us the highest F-measure (0.74) as compared to translated lexicons and other Arabic lexicons.

REFERENCES

Abdaoui, A., Azé, J., Bringay, S., & Poncelet, P. (2016). FEEL: a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 1-23. doi:10.1007/s10579-016-9364-5

Bermingham, A., & Smeaton, A. F. (2010). *Classifying sentiment in microblogs: is brevity an advantage?* Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2016). Building a Twitter opinion lexicon from automatically-annotated tweets. *Knowledge-Based Systems, 108*, 65-78. doi:10.1016/j.knosys.2016.05.018

Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems, 94*, 65-76.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research, 50*, 723-762.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., . . . Zavarella, V. (2012). Creating sentiment dictionaries via triangulation. *Decision Support Systems, 53*(4), 689-694. doi:http://dx.doi.org/10.1016/j.dss.2012.05.029

Wu, F. Z., Huang, Y. F., Song, Y. Q., & Liu, S. X. (2016). Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decision Support Systems, 87*, 39-49. doi:10.1016/j.dss.2016.04.007

Evaluation Datasets for Research Paper Recommendation Systems

Khalid Haruna[1,2] and Maizatul Akmar Ismail[1]

[1] *Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia*
[2]*Department of Computer Science, Faculty of Computer Science & Information Technology, Bayero University, Kano, Nigeria*

Corresponding Emails: kharuna.cs@buk.edu.ng, maizatul@um.edu.my

Keywords: *Research Paper, Evaluation Datasets, Recommendation Systems*

INTRODUCTION

To provide recommendations to users, merely having a recommendation algorithm is not sufficient. The algorithm needs data about users to model their preferences and ultimately provide them with a customised experience (Ekstrand, Riedl, & Konstan, 2011). However, there is no appropriate dataset available for the overall evaluation of research paper recommendation algorithms (Pan & Li, 2010). Notwithstanding, many digital libraries and search engines have made available their data sets which can be downloaded freely or by requesting access to help researchers better explore academic society. In this paper, we perform an extensive survey to identify the most used evaluation datasets from the literature.

METHODOLOGY

In an attempt to perform an exhaustive search, the most accurate and reliable bibliographic databases that cover the most important journal articles and conference proceedings are identified. These databases are Science Direct, ACM Digital Library, Springer Link, Web of Science, IEEE Xplore, Scopus, DBLP, Direct Open Access Journals (DOAJ), ProQuest, CiteSeeX, arXiv, Microsoft Academic Search, and Google Scholar portal. A Boolean search criterion was then used to search the bibliographical databases. "(Title ((paper OR article OR scholarly OR citation) AND (recommender OR recommendation)) OR abstract ((paper OR article OR scholarly OR citation) AND (recommender OR recommendation)). Each of the extracted papers was then prudently reviewed to identify the dataset they used in the evaluation process.

FINDINGS

Based on the conducted literature, the most used evaluation dataset for validating the effectiveness of research paper recommendation algorithms is summarised in Table 1.

Table 1 Most used Evaluation Datasets for Research Paper Recommendation Systems

Dataset	Description	Type	URL
ACM portal	The ACM Digital Library is a research, discovery and networking platform that contained; (a) Full-Text Collection of all ACM publications, including journals, conference proceedings, technical magazines, newsletters and books. (b) A collection of curated and hosted full-text publications from selected publishers. (c) The ACM Guide to Computing Literature, a comprehensive bibliographic database focused exclusively on the field of computing. (d)A richly interlinked set of connections among authors, works, institutions, and specialised communities.	Real Data	https://dl.acm.org/
Citeseer	Citeseer is a public search engine and digital library for scientific and academic papers. The operation of CiteSeer is relatively simple. Given a set of broad topic keywords, CiteSeer uses Web search engines and heuristics to locate and download papers that are potentially relevant to the user's topic. The downloaded papers are parsed to extract semantic features, including citations and word frequency information. This information is then stored in a database which the user can search by	Real Data	https://www.ebi.ac.uk/training/online/glossary/citeseer#

	keyword, or use citation links to find related papers. The agent can also automatically find papers similar to a paper of interest using semantic feature information.		
CiteULike	CiteULike is a free service for managing and discovering scholarly references	Real Data	http://www.citeulike.org/
CiteSeerX	CiteSeerX is an evolving scientific literature digital library and search engine that has focused primarily on the literature in computer and information science. CiteSeerX aims to improve the dissemination of scientific literature and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness in the access of scientific and scholarly knowledge. Rather than creating just another digital library, CiteSeerX attempts to provide resources such as algorithms, data, metadata, services, techniques, and software that can be used to promote other digital libraries. CiteSeerX has developed new methods and algorithms to index PostScript and PDF research articles on the Web.	Real Data	http://csxstatic.ist.psu.edu/about
Sugiyama and Kan (2015)	Sugiyama and Kan (2015) Composed experimental dataset that consists; (a) Feature vectors of candidate papers to recommend (from ACM DL). (b) Citation and reference information about each candidate paper to recommend. (c) Research interests of 50 researchers. (d) Feature vectors generated from each researcher's published papers in DBLP list and (e) Paper IDs relevant to each researcher's interests	Real Data	https://www.comp.nus.edu.sg/~sugiyama/Dataset2.html
Corpus-ACL Anthology Network (AAN)	The ACL Anthology Network was built from the original PDF files available from the ACL Anthology. It contains information regarding all of the papers included in the many ACL venues, including paper citation, author citation, and author collaboration.	Real Data	http://clair.eecs.umich.edu/aan
Microsoft Academic Search (MAS)	Microsoft Academic is a public search engine for academic publications and literature, developed by Microsoft Research. The tool features an entirely new data structure and search engine using semantic search technologies.	Real Data	http://academic.research.microsoft.com/
DBLP	DBLP is a digital library that provides open bibliographic information on major computer science journals and proceedings.	Real Data	http://dblp.uni-trier.de/
TREC-KBA	TREC-Knowledge Base Acceleration is an open evaluation dataset that provides training data in the form of ground truth, which contains information of Text Retrieval Conferences (TREC) of 2012, 2013 and 2014.	Real Data	http://trec.nist.gov/data/kba.html

CONCLUSION

Research paper recommendation algorithms need a dataset to operate, which must be obtained from users to model their preferences and ultimately provide them with relevant suggestions. This paper performs an extensive survey to identify the most used dataset in validating the effectiveness of research paper recommendation algorithms. The identified datasets can either be downloaded freely or by requesting access. One important observation is that all of the identified data sets are real data and none is synthetic, which means the results obtained from each approach is likely to be the same when applied in the real world.

ACKNOWLEDGEMENT

We warmly thank our colleagues for their valuable support and assistance. This research is supported by UM Research Grant No. RP059B-17SBS.

REFERENCES

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction, 4*(2), 81-173.

Pan, C., & Li, W. (2010). *Research paper recommendation with topic analysis.* Paper presented at the Computer Design and Applications (ICCDA), 2010 International Conference on.

Sugiyama, K., & Kan, M.-Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries, 16*(2), 91-109.

Support System for Parents with Dyslexia Children: An Overview

Athira Amira Abd Rauf and Maizatul Akmar Ismail

Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia

Corresponding Emails: athiramira94@gmail.com, maizatul@um.edu.my

Keywords: *Support, Parent, Dyslexia*

INTRODUCTION

"Dyslexia" is a life-long specific, isolated impairment of reading and spelling. (Schulte-Körne, 2010). And it is a life-long disability with its symptoms vary from person to person depends on the person's life age and environment but with appropriate intervention, it can produce a significant result (Skiada, Soroniati, Gardeli, & Zissis, 2014). Therefore, parents with special disabilities children with limited support and resources (Papageorgiou & Kalyva, 2010) tends to experience large number of challenges (Heiman & Berger, 2008) and a higher stress level (Papageorgiou & Kalyva, 2010). Thus, the aim for this research is to identify and understand the type of supports needed by parents in raising their dyslexic child. Furthermore, the novelty of this research is the state of art of Dyslexia issues and important components for model development, the model of support system and the support system tool for Dyslexia community and can help promote awareness and improve support system for Dyslexia community.

PROBLEM STATEMENT

Dyslexia is a life-long disability with its symptoms varies from person to person depends on the person's life age and environment but with appropriate intervention, it can produce a significant result (Skiada, Soroniati, Gardeli, & Zissis, 2014). Based on observation, most research papers were more targeted toward the development techniques, challenges and supports of dyslexic children rather than the supports and methods needed by parents in raising their dyslexic child. Other than that, research papers that discuss about the parents support model tend to discuss only one type of support like group support (Papageorgiou & Kalyva, 2010) or family support (Heiman & Berger, 2008). According to Wang & Michaels (2009) study shows that 27.35% of parents have trouble obtain specialist like doctors and teachers information from the web or social media. Meanwhile, 64.5% of parents have trouble obtaining the latest information and news on the development and treatment in the Dyslexia community (Papageorgiou & Kalyva, 2010). Therefore, due to the difficulties in finding necessary information and support needed by parents with dyslexic children it can not only affect the dyslexic child, but it can also affect the parents themselves.

METHODOLOGY

This research will be using the quantitative descriptive research design. Currently, the research is focusing on obtaining the types of supports that parents need in raising their children with dyslexia from doing literature review by reading previous research papers and conducting online survey and focus groups to obtain the parents opinions regarding the supports and awareness methods.

FINDINGS

The important factors in raising a dyslexia child is the parent's knowledge and support regarding the disability (Papageorgiou & Kalyva, 2010) by collecting information from a formal support services like professional therapist (Meadan, Halle, & Ebata, 2010) and from the Internet or social media where they can communicate with other parents support groups (Paquette, 2013).

Based on previous studies, it shows there are various types of supports that parents need to raise their dyslexic children such as social support, organizational supports (Heiman & Berger, 2008), financial support and emotional supports (Cen & Aytac, 2017). Social support from either family member or friends (Papageorgiou & Kalyva, 2010) is important because it can help parents to improve their mental health by reducing their stress levels and reduce the parents depression risks (Benson, 2006; Meadan, Halle, & Ebata, 2010).

Meanwhile, financial support like additional education aids and health insurance coverage is also an important support that parents need in raising their dyslexia children (Cen & Aytac, 2017). Other than that, according to Papageorgiou and Kalyva (2010) research shows that parents that participate in a peer-to-peer support groups tend to feels less isolated and stressful due to the child, and it can also help parents to learn and receive advice about new methods of handling their dyslexic child. According to Wang & Michaels (2009) research, it shows that 27.35% of parents demands more professionals like doctors, teachers and therapist that are specialist in dyslexia in a government schools or hospitals. Lastly, 22% of parents wants the government to increase the society awarness about dyslexia and increase their governmental supports level and funding (Wang & Michaels, 2009).

CONCLUSION

In conclusion, "Dyslexia" is a life-term disability that is genetic and hereditary, and which can affect both children and adults. The most important treatment for dyslexic children is the parents' knowledge about dyslexia and its impact on their children. Parents need to be made aware of supports and resources that they need to take care their dyslexic children. At this stage, we are currently examined the type of supports parents need and ways parents are aware regarding dyslexia. Our next line of research is to propose a supportive model that will increase the parent awareness of the factors leading to children learning difficulties, thus providing a support system for parents with dyslexic children.

ACKNOWLEDGEMENT

We warmly thank our colleagues for their valuable support and assistance. This research is supported by UM Research Grant No. RP059B-17SBS.

REFERENCES

Alias, N. A., & Dahlan, A. (2015). Enduring Difficulties: The Challenges of Mothers in Raising Children with Dyslexia. *Procedia - Social and Behavioral Sciences*, 107-114.

Cen, S., & Aytac, B. (2017). Ecocultural Perspective in Learning Disability: Family Support Resources, Values, Child Problem Behaviors. *Learning Disability Quarterly 2017*, 114 - 127.

Heiman, T., & Berger, O. (2008). Parents of children with Asperger syndrome or with learning disabilities: Family environment and social support. *Research in Developmental Disabilities*, 289-300.

Meadan, H., Halle, J. W., & Ebata, A. T. (2010). Families With Children Who Have Autism Spectrum Disorders: Stress and Support. *Exceptional Children, 77*(1), 7-36.

Papageorgiou, V., & Kalyva, E. (2010). Self-reported needs and expectations of parents of children with autism spectrum disorders who participate in support groups. *Research in Autism Spectrum Disorders*, 653-660.

Paquette, H. (2013). *Social Media as a Marketing Tool: A Literature Review.* University of Rhode Island.

Schulte-Körne, G. (2010). The Prevention, Diagnosis, and Treatment of Dyslexia. *Deutsches Ärzteblatt International*, 718-727.

Wang, P., & Michaels, C. A. (2009). Chinese Families of Children With Severe Disabilities: Family Needs and Available Support. *Research & Practice for Persons with Severe Disabilities*, 21-32.

Multiple Feature Enhanced Cyberbullying Detection Model

Shahzaib Khan, Vimala Balakrishnan and Ng Koi Yee
Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Malaysia

Corresponding Emails: shahzaib198@gmail.com, vimala.balakrishnan@um.edu.my

INTRODUCTION

Cyberbullying refers to "any behavior performed through electronic media by individuals or groups of individuals that repeatedly communicate hostile or aggressive messages intended to inflict harm or discomfort on others" (Tokunaga, 2010). It is also described as a deliberate repetitive attempt to harm a victim via mobile phones, e-mail, Internet chats, social media and personal blogs (Hinduja & Patchin, 2008) with threatening messages and spreading malicious rumors online. It has now emerged as the main problem in social media where users are able to do and say anything anonymously, resulting in increased life-threatening incidents amongst younger victims of cyberbullying.

A number of researches have been done on cyberbullying (Park et al., 2014; Popović-Ćitić et al., 2011), mostly related to its prevalence from the social perspective. However, little attention has been given to the online detection mechanisms, except for a few (Hosseinmardi et al., 2015; Patch, 2015; Chatzakou et al., 2017). These studies also lack in the study of content, users and network features.

The study aims to extract the extended features from the tweets, network, and users profile and propose an effective model to detect cyberbullying activities on Twitter. The research extends the state of the detection methods by proposing a methodology for large-scale analysis and extraction of content, user, and network based features. The study also examines the attributes which will further distinguish Twitter user behaviors. Finally, the study evaluates the effectiveness of the detection model based on the result from the trained classifier.

METHODOLOGY

The detection of cyberbullying behavior on Twitter includes the following steps:

Data Collection: The data were collected via Twitter's Streaming API using a labeled set of Twitter ids provided by Chatzakou et al. (2017), resulting in 9484 records.

Preprocessing: Preprocessing of data to remove the stop words, URLs, etc. and normalization.

Feature Extraction: Text-based-features (e.g. text length, keywords, etc.), and user-based features (e.g. followers count, friends count, statuses count, etc.).

Classification: The labeled dataset and features were used in multiple classifiers (Naïve Bayes, J48, Random forest) to gauge the best classifier.

Evaluation: Experimental analysis was done to further conclude the efficiency of model in cyberbullying detection. AUC is used as our main performance measure because of its high robustness.

FINDINGS

An evaluation demonstrated that Random forest outperformed the other classifiers and achieved an overall accuracy of 90.7% (97.7% AUC) with the most significant features. The accuracy, further enhanced as a dimensionality reduction algorithm was applied and the same classifier was able to achieve overall accuracy of 91.7% (99% AUC) with the least set of features. This indicates that network features are good in detecting cyberbullying compared to user-based features in the detection of cyberbullying especially with the dimensionality reduction algorithm applied.

CONCLUSION

This research developed a model for detecting cyberbullying on Twitter. It advances the cyberbullying detection method beyond textual analysis by including users and network-based features. The model achieved an overall accuracy of 91.7% with AUC 99%. Compared to existing studies, the model has significantly improved overall accuracy with the least set of features. In future, a similar approach can be applied over more fine-grained data on human ehaviour to detect cyberbullying in different scenarios, thus paving the way for a safer environment for victims.

REFERENCES

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433-443. doi:https://doi.org/10.1016/j.chb.2016.05.051

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. Paper presented at the Proceedings of the 2017 ACM on Web Science Conference.

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. Deviant behavior, 29(2), 129-156.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. Paper presented at the International Conference on Social Informatics.

Park, S., Na, E.-Y., & Kim, E.-m. (2014). The relationship between online activities, netiquette and cyberbullying. Children and youth services review, 42, 74-81.

Patch, J. A. (2015). Detecting bullying on Twitter using emotion lexicons. University of Georgia, Popović-Ćitić, B., Djurić, S., & Cvetković, V. (2011). The prevalence of cyberbullying among adolescents:

A case study of middle schools in Serbia. School psychology international, 32(4), 412-424. Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on

cyberbullying victimization. Computers in Human Behavior, 26(3), 277-287.

The Impact of Letter Repetition on Sentiment Strength

Rayventhiran Visvalingam
Ipsos Sdn Bhd

Corresponding Email: rayven_sri@yahoo.com

Sentiment polarity calculation is a method to gauge the strength of a sentiment extracted from a text. Many tools have been developed with their respective scoring mechanism in order to produce an effective sentiment score. One such tool is the Semantic Orientation Calculator (SO-CAL), a lexicon-based tool that is incorporated with important features such as intensifiers, negation, etc. to calculate the sentiment polarity of a text. However, this tool has its limitation in analyzing misspelled words, especially in repeated letters or characters that may lead to sentiment inaccuracy. The accuracy of SO-CAL is affected when processing social media text that mostly contains misspelled words. Studies have indicated each repeated letter to indicate a stronger sentiment. When a user is over excited or deeply disturbed by an event, it triggers them to vent out such emotions via letter repetition. Hence ignoring such repetition as merely human error would cause an inaccurate sentiment score that would skew true sentiment causing businesses to misinterpret their customers' needs.

The study aims to enhance the scoring mechanism by considering misspelled words, especially words that contain repeated letters. The tool, aptly named as Sentiment Intensity Calculator (SentI-Cal) was tested based on Facebook posts from two major airline industries in Malaysia, Airline A and B. Three important phases were involved in the development of SentI-Cal: data collection, data cleaning and data analysis. Data collection was performed with the aid of Facebook Graph API to collect three months' posts from the both airlines. This was followed by the data cleaning phase, in which noise was removed, leaving only text that contains alphabets and exclamation marks. Improvement was made on the scoring mechanism and incorporated in SentI-Cal with the features that can process misspelled word, and also other improved features such as negation and inclusion of exclamation mark.

Comparisons were made between SentI-Cal and SO-CAL using evaluation metrics such as accuracy, recall, precision and F1-score, with the reference of human experts. Results show that SentI-Cal achieved a higher accuracy (90.7%) than SO-CAL (58.33%.). Overall results also show that Airline A achieved a high positive score than Airline B. This concludes processing misspelled word is an important process in social media sentiment analysis. This is further proved with the reference to the case study, where a conclusion was formed as Airline A providing a better service than Airline B.

Data Science, Sentiment Analysis and the Impact of Free Speech on Social Media

Terence Fernandez
Analyst, Spatial Media LLC, USA

Corresponding Email: conciseversion@gmail.com

Data Science, has evolved from the seemingly sort of thing mysteriously practised by mathematicians, statisticians and economist, with their extent of deep organic knowledge to fathom its power. Today, data isn't a mystery anymore. Data Science is endemic to modern life and it has permeated every aspect of it. We have traditionally assumed the term data to be information that is structured, formatted in a large database of sorts. However, a greater part of our world of information isn't just structured. It is unstructured, text-heavy, language diversified, location based, and contain differing data narratives.

Sentiment analysis is an area of research that involves the decrypting of unstructured data for the purpose of establishing the attitude of an author with respect to a subject matter (Liu, 2012). Sentiment analysis methods help classify and understand users' feelings on a topic of interest. The demand to extract and decipher information from amorphous data has seen the field of sentiment analysis expanding from product reviews to disaster management, and of late, electoral and political analysis (Ahmed et al., 2014; 2018; Adedoyin-Olowe et al., 2016), among others. In this abstract, we provide a sentiment analysis of Twitter discussions on the recent 14th General Election (GE14) in Malaysia, which resulted in the long-standing Barisan Nasional government being ousted by the opposition party. This abstract presents specific findings based on a preliminary analysis conducted on randomly selected tweets.

A data set consisting of 175 343 Twitter messages were extracted using the hashtag #GE14, targeting tweets posted on the day of election. The resulting data set is multi-dimensional, including user location, as well as information on the user's emotions expressed in the content of the tweets. The extracted tweets were first screened, all re-tweets removed, along with tweets containing only URLS. The final data set contained 65 370 tweets, comprising both Malay and English language tweets. Content-analysis approach was applied whereby 2500 tweets were randomly selected, and manually analyzed by the researcher, assisted by two other linguistic experts. The analysis was mainly focused on tweet categories (i.e. information sharing, opinion expression, information seeking, insinuation and others). In addition, the sentiments of the tweets were also determined.

Of the 2500 tweets analyzed, Malay was the preferred language by majority of the users (63.7%). As for the categories, majority of the tweets were for opinion expression (43.5%) with almost equal positive and negative sentiments (i.e. 56% vs. 44%), information sharing (38.7%; positive), followed by insinuation (9.4%; negative), information seeking (5%: neutral) whilst the rest were categorized as Others (3.4%; neutral). The results show that tweets categorized as insinuation, although small in percentage, were entirely negative in nature. Such data may be deemed important and require further investigation as to the source and reasoning behind such negative insinuation. It is important to note that this is a study of tweets only from the day of election, therefore, further analysis of tweets from a time period before and after the election will reveal in-depth information that may be beneficial to various studies essential to political analysis.

REFERENCES

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., & Gomes, J. B. (2016). A rule dynamics approach to event detection in Twitter with its application to sports and politics. *Expert Systems with Applications*, *55*, 351-360

Ahmed, S., Jaidka, K. & Cho, J. (2018) Do birds of different feather flock together? Analyzing the political use of social media through a language-based approach in a multilingual context, Computers in Human Behavior, 86, 299-310

Ahmed, S., Jaidka, K. & Cho, J. (2014) The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties, Telematics and Informatics, 33 (4) 1071– 1087

Mapping Human Emotions using Keyword Based Emotion Filter and String Vectors

Wandeep Kaur[1], Amir Javed[2], Vimala Balakrishnan[1], Pete Burnap[2], Hajar Abdul Rahim[2]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[2]Cardiff University, School of Computer Science and Informatics

[3]School of Humanities University Sains Malaysia

Corresponding Emails: wandeep@gmail.com, vimala.balakrishnan@um.edu.my

INTRODUCTION

Traditionally, documents are encoded using numerical vectors where words mined from a corpus are carefully chosen as features by a criterion or a combination of criteria (Diab & El Hindi, 2017). However, the use of typical feature selection methods such as TF-IDF, information gain and chi square pose two major concerns; huge dimensionality and sparse distribution. Huge dimensionality refers to the number of features vigorously needed to represent documents into numerical vectors (Jo, 2016). When it comes to classifying large amounts of textual data, the number of features extracted from within each input can reach several ten thousand and despite the availability of advanced feature selection methods, these numbers can only be reduced to several hundred (Shalin & Prasadh, 2016). In the experimentations done for this study it was found that large number of features tend to confuse the algorithm thus resulting in inaccurate classification of text. Sparse distribution on the other hand refers to the representation of zero value in the numerical vector representing document where zero value takes up more than 90% of each numerical vector representation causing the feature to have very limited coverage in a given corpus leading the computation of the inner product to be very friable (Jo, 2016).

This study looks to improve on existing machine-learning algorithm to their string vector-based versions using a keyword-based emotion filter. A similarity matrix is automatically built from the corpus and the Multinomial Naïve Bayes algorithm (MNB) is modified into string-based version for classification purpose. The similarity matrix works with using bag of words concept where the n-gram of words around the emotion word is taken into consideration to build a string of words that eventually categorizes the said emotion.

METHODOLOGY

The corpus used for this study is posts extracted from a Facebook diabetes community and the extracted posts are annotated where the inter-coder reliability is calculated at 89% and the Krippendorf alpha has also been calculated (Krippendorff, 2004). However, only a total of 800 posts were used for the purpose of this experimentation.

A keyword-based emotion filter was used to improve the overall f-measure of the classifier. The aim of the filter was to identify posts that contain some form of emotion based on keyword extracted from said posts. Each post goes through the standard pre-processing of stop word removal, stemming and lemmatization before it is passed into the filter. The filter then detects eight primary emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) as defined by Plutchik (2003). The filter uses multiple dictionaries such as Wordnet Affect Lexicon (Strapparava & Valitutti, 2004) and NRC Emotion Lexicon (Mohammad & Turney, 2013) containing words and emotions associated with each word.

Once the emotions have been identified, the post is then passed through a Multinomial Naïve Bayes algorithm that converts each post into a string of vectors to map the words around the word identified as an emotion using bag of words concept. From here, the classifier works to classify said post according to Plutchik (2003) emotion.

FINDINGS

The results shown in Table 1 are a comparison of with filter applied and without filter applied. The standard evaluation measure of accuracy, recall and f-measure was calculated. However, for the purpose of this abstract, only one emotion results will be discussed.

It has been found that with the use of the filter to identify emotion and have the classifier then correctly classify the emotion, the f-measure results were recorded at 0.884 compared to 0.686 without the filter mechanism. This is due to the hierarchical classification of first identifying emotions within a post using WordNet Affect Lexicon and NRC Emotion Lexicon dictionaries before classifying the post according to emotion using Multinomial Naïve Bayes algorithm using string vectors. The significant jump is also credited to the reduce dimensionality of space within n-gram words that associate words around the emotion identified word.

Table 1: Findings of with filter applied against without filter

Emotion	With Filter			Without Filter		
	Accuracy	Recall	F-measure	Accuracy	Recall	F-measure
Trust	0.805	0.980	0.884	0.627	0.758	0.686

CONCLUSION

In this paper, keyword-based emotion filter is used to first detect the emotion found within the said Facebook posts. This allows the algorithm to easily identify the emotion words and associate the words around the emotion word using bag of words concept to accurately classify the posts to its correct emotion. This can be seen through the results displayed in Table 1 above.

ACKNOWEDGEMENT

The authors would like to thank and acknowledge the support provided by University of Malaya, under research grant reference number: UMRG RP059C 17SBS.

REFERENCE

Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing, 54*, 183-199. doi:https://doi.org/10.1016/j.asoc.2016.12.043

Jo, T. (2016). *Encoding Words into String Vectors for Word Categorization.* Paper presented at the Proceedings on the International Conference on Artificial Intelligence (ICAI).

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research, 30*(3), 411-433.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436-465.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*: American Psychological Association.

Shalin, L. V. A., & Prasadh, K. (2016). An Effective Multi-clustering Anonymization Approach Using Discrete Component Task for Non Binary High Dimensional Data Spaces. *Procedia Technology, 25*, 208-215. doi:https://doi.org/10.1016/j.protcy.2016.08.099

Strapparava, C., & Valitutti, A. (2004). *Wordnet affect: an affective extension of wordnet.* Paper presented at the Lrec.

Internet of Things and Block Chain - A Promising Digital Fusion

Terence Fernandez
Analyst, Spatial Media Ltd., USA

Corresponding Email: conciseversion@gmail.com

INTRODUCTION

The terms blockchain and Internet of Things (IoT) have created so much of hype not only within the realm of technology, but in the business circle as well. IoT describes the on-going proliferation of always-online, data-gathering devices into our work and personal lives, with a prediction of having more than 20 billion connected devices by 2020. With such a huge number of connected devices, it is no surprise that identifying, connecting, securing, and managing these become complicated. On the other hand, a blockchain is an encrypted, fully-transparent, permission-less, distributed computer ledger designed to allow the creation of tamper-proof, real-time records directly between different participants. The blockchain basically runs on trust protocols that mean nobody can alter any records. With such attractive features, the idea of combining IoT and blockchain seems only logical. In fact, technology pundits estimate IoT and blockchain to be integrated within the next five years. However, not every data collected from IoT devices are suitable for blockchain.

AGRICULTURE

IoT helps by optimizing supply chains and deliver smart logistics by tracking assets from a farmer's field to the consumers through a web of connected sensors (i.e. food supply chain). An integration with blockchain will guarantee food safety, for example. Filament, a start-up company is already using IoT tools in both manufacturing and agriculture, along with device identification and communication secured with a Bitcoin blockchain.

SMART HOMES

Much of the data generated by IoT is highly personal, including smart home devices that have access to personal details of our daily lives and routines. A classic example would be Telstra, an Australian telecommunication company that uses private blockchains to minimize verification time, with no tamper resistance. User biometric information are added to their blockchain protocol, and thus providing an increased security feature to those interacting with those devices.

The merging of IoT and blockchain is not only limited with the two aforementioned domains, in fact, it can be extended to other fields including healthcare, shipping industry, entertainment industry, finance, and even voting mechanisms. The benefits of the digital fusion between IoT and blockchain in undeniable, however, it comes with several challenges as well. These technologies are considered as budding technologies, proper standards need to be defined. Blockchain for example, is limited in terms of its scalability and confidentiality, whilst IoT technology needs to prove that its infrastructure is resilient and efficient. Finally, not all data gathered from IoT devices are suitable for blockchain.

Educational Analytics – Opportunities, Challenges and Directions

Tharmaraj Kandasamy
Real Estate Vertical
TM One @ TM Berhad

Corresponding Email: k_tharmaraj@yahoo.com

INTRODUCTION

Data analytics refers to methods and tools for analyzing large sets of data from diverse sources aiming to support and improve decision making. Hence, data has become a strategic asset, and coupled with increased computing capacity and speed, data analytics have changed the way businesses make decisions. Although data analytics includes technologies applied in financial, business and health systems, it has only recently been considered in the context of education, both in higher education and school education.

OPPORTUNITIES

The educational institutions have long amassed data, such as students' registration, grades, attendance, and the like. With large volumes of data, variety and high velocity in speed of data, universities, for example can convert and diagnose these data into information and obtain an insight to make the right decisions. As a matter of fact, with the adoption of technology in more the education sector, there are clearly a lot of opportunities for better data gathering and analysis in education.

CHALLENGES

However, despite its importance, the current level of use of educational data is still limited, mainly due to a number of barriers. For example, privacy issues. Legal and privacy concerns vary between countries, nevertheless data are often considered private, and this might hinder institutions to fully explore and utilize the data available. Other barriers include lack of technical capabilities. In Malaysia for example, data science and analytics is an emerging trend, hence skills involved in educational data analytics are lacking. Low quality of educational data can be a potential issue as well, considering that quality of data in hand will significantly affect the computational outcome, which eventually effect the decision making process.

DIRECTIONS

Nevertheless, as data science and analytics are growing steadily in various fields, it it important, and also timely to build a culture of interpreting statistics and thinking of numbers among people. As a matter of fact, this is already embodied in the local education curriculum, however, institutions should emphasize more on data analytics, especially in exploring ways and tools that can be used, such as RapidMiner, SAS Enterprise, Tablaeu etc.

An Applied Machine Learning Approach to Predict the Treatment Effectiveness of Medicinal Plants

Vala Ali Rohani[1], Sharala Axryd[1], Ami Fazlin Syed Mohamed[2], Tan Yee Sern[3] and Yaszrina binti Mohd Yassin[4]

[1] Department of Research and Development, ASEAN Data Analytics eXchange (ADAX)
[2] Herbal Medicine Research Centre Institute for Medical Research (IMR)
[3] Department of Data Science, The Center of Applied Data Science (theCADS)
[4] Data Economy Division, Malaysia Digital Economy Corp (MDEC)

Corresponding Email: vala@adax.asia

INTRODUCTION

In herbal product development, preclinical studies are conducted to determine the suitability of the product before the progression into clinical study. The steps in the preclinical study include product standardization, proof of efficacy and evaluation of toxicity. Standardization is a process that involves the identification of the plant by botanical and phytochemical methods as well as ensuring that certain standard chemical markers are present in the extract (LEAF, 2015). This process would ensure that herbal products in the same category, produce similar efficacy and side effects.

Currently, the screening of the medicinal plants' efficacy is based on their reported traditional use (Burkill, 2015). For example, plants reported to be used for fever may have antibacterial, antimalarial or anticancer properties. The medicinal plants are unique as they may produce similar properties and yet be different. The properties of the medicinal plants are determined by its composition. Each would be based on similar components for example carbohydrates, alkaloids, terpenoids, flavonoids and others (LEAF, 2015). These compounds would exist in different proportions for the different plants. Different types of extractions processes will yield different types of extracts and the fingerprints can be identified. The phytochemicals from specific types of extract can be fractionated and further processing will enable specific compounds to be identified (LEAF, 2015). The complex chemical composition determines the biological properties i.e. the efficacy and toxicity of the plants. Preclinically, both efficacy and toxicity are evaluated by specific in vitro and in vivo models. The accumulative cost of screening and carrying out the in vitro and in vivo studies are moderate to high especially with highly specific models.

So, with the aim of reducing the cost of screening process, in this paper we propose a machine learning based solution to predict the efficacy of herbal plants in treatment of Malaria and some types of cancer.

METHODOLOGY

In this research we collected 66 records of analysis results for 16 types of herbal plants, based on a two-stage screening approach.

In first stage screening, the extracts were tested against Plasmodium falciparum K1 (a parasite that cause malaria in human) using HRPII assay (Histidine-Rich Protein II Enzyme-Linked Immunosorbent). The result of the first stage is indicated by the EC50 value, a dose at which 50% of the maximum effect is produced or the concentration of drug at which the drug is half-maximally effective. This is commonly used as a measure of a drug's potency and the unit is μg/mL. When the EC50 showed weak activity (> 15.70 μg/mL), the extract will not be tested in the second stage screening.

In the second stage screening, the extracts were tested against MDBK cells using MTT assay (a cell viability assay which depends upon a mitochondrial dehydrogenase acting upon MTT to produce dark-blue formazan from MTT (3-(4,5-dimethylthiazol-2-yl)- 2,5-diphenyltetrazolium bromide). Madin-Darby bovine kidney (MDBK) cells is a mammalian cells used to test for cytotoxicity. The result of the second stage is also indicated by EC50 value (μg/mL).

In analytics part of this research project, after cleansing the datasets and identifying the predictors and random variable, we applied 6 different machine learning algorithms, including Decision Trees (Rokach & Maimon, 2008), Logistic Regression (Hosmer Jr, Lemeshow, & Sturdivant, 2013), Neural Networks(Rojas, 2013), Random Forest (Boulesteix, Janitza, Kruppa, & König, 2012), AdaBoost (Ying, Qi-Guang, Jia-Chen, & Lin, 2013), and Gradient Boosting Machines(Natekin & Knoll, 2013) to build the highest accurate model for predicting the most effective herbal plants in treatment of studied cancers and Malaria. Also, to mitigate the small size of collected data, we applied bootstrapping technique (Ratner, 2017) with 200 numbers of iterations. The details of experiments and results of each model are discussed in the next section.

EXPERIMENTS AND RESULTS

Based on the dataset provided by Institute of Medical Research in Malaysia comprises the results of antiplasmodial and cytotoxicity tests on hebal plants, we conducted data cleansing and exploratory analysis to make it prepare for next phase of predictive modeling. Afterwards, some properties of medicinal plants, such as plant name, family, part, and type of extracts were selected as independent variables to predict the remark levels in treatment of cancers and Malaria.

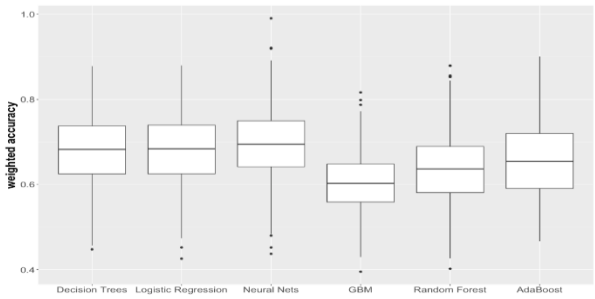


Figure 1. Weighted accuracy of ML models for cancers treatment

To find the best prediction model with the highest accuracy rate, six different machine learning algorithms of classification were applied. According to Figure 1 to compare the weighted accuracy of Machine Learning models to predict the effectiveness of medicinal plants in cancers treatment, Neural Network model yield the best performance of 72% followed closely by Decision Trees with 68% of accuracy.

For Malaria treatment, we recorded the better performance of 75% for Gradient Boosting Machine and 70% for Decision Trees prediction models (Figure 2). This was because of having more numerical values in two-stage test results for Malaria related experiments, which enabled us to include them in the models as predictors.

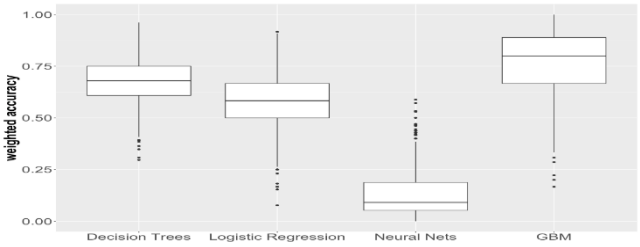


Figure 2. Weighted accuracy of ML models for Malaria treatment

CONCLUSION

In this study, we applied six different machine learning models to predict of the efficacy of 16 studied medicinal plants in treatment of cancers and Malaria. The generated machine learning models in this study, can be used in pharmacology industries to decrease the cost of expensive antiplasmodial and cytotoxicity tests needed in predicting the effectiveness of plants for treatment of different cancers and Malaria.

For future works, it's suggested to collect more samples to achieve even higher performance of predictions.

ACKNOWLEDGEMENT

This research work is supported by ASEAN Data Analytics eXchange (ADAX), MDEC, and the institute of medical research (IMR).

REFERENCES

Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 493-507.
Burkill, I. H. (2015). A dictionary of the economic products of the Malay Peninsula.
Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.
LEAF, M. C. (2015). MALAYSIAN HERBAL MONOGRAPH.
Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7*, 21.
Ratner, B. (2017). *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data*: Chapman and Hall/CRC.
Rojas, R. (2013). *Neural networks: a systematic introduction*: Springer Science & Business Media.
Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69): World scientific.
Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica, 39*(6), 745-758.

Distress Level Detection using Emotion Detection Techniques – A Conceptual Framework

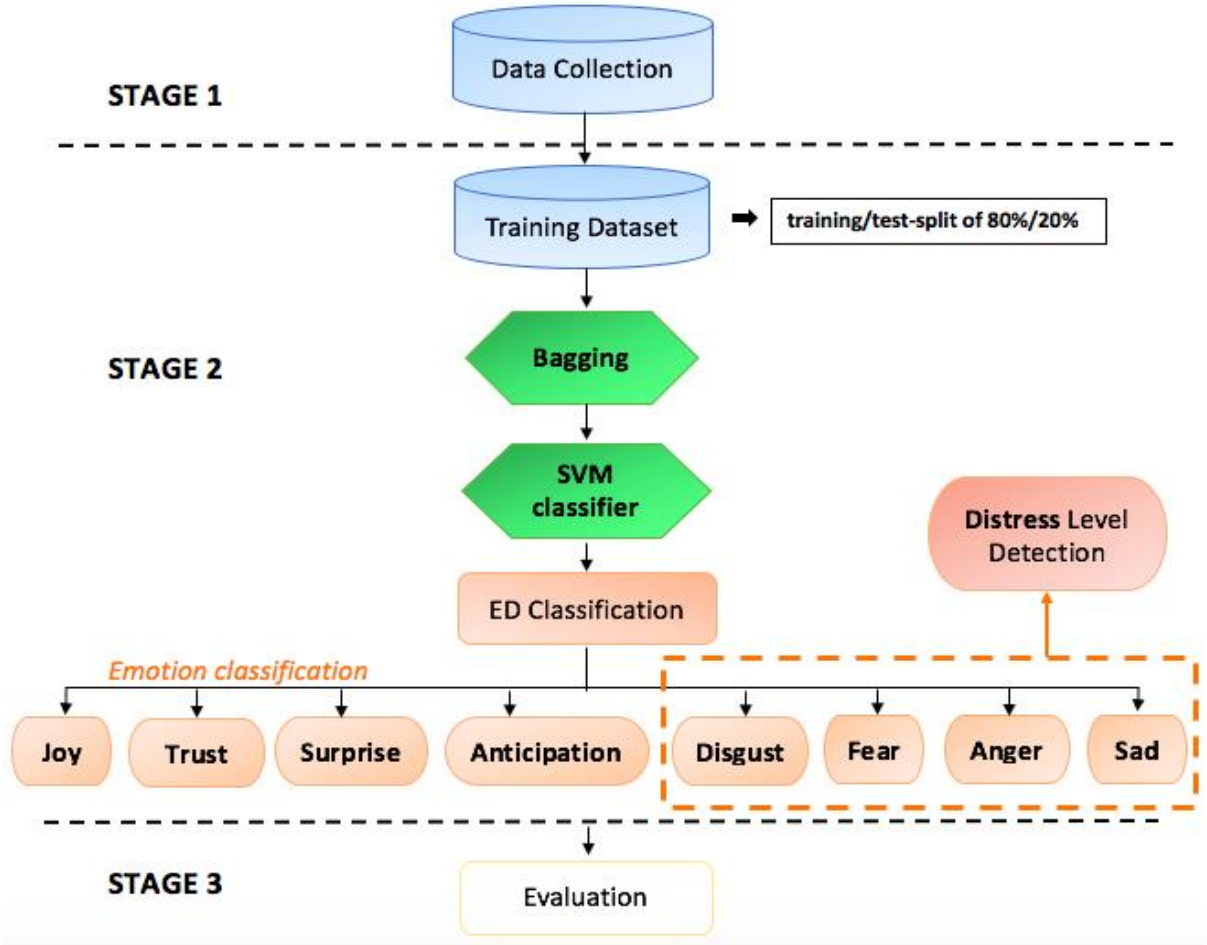
Marian Cynthia Martin
i-Tech Network Solutions

Emotion is any conscious experience intertwined with personality, mood, temperament and motivation. Its plays an important role influencing overall human behavior where reasoning, decision making and interaction are affected. Emotion Detection (ED) is an opinion mining task concerns the computational study of natural language expressions in recognizing various emotions from text (Liu & Zhang, 2012). ED has proven to have wide applications ranging from building nuanced virtual assistants that cater for the emotions of their users to detecting the emotions of social media users in order to understand their mental and/or physical health. With the explosive expansion of social media has produced massive volume of data available in the form of text, thus resulting opinion mining to become an active domain to extract knowledge from this text. ED aims to reflect the content of the text into various emotions such as joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Generally, ED can be performed on text, emoticons, reactions, facial expressions and audios though the common ones are text and audio.

This paper presents the findings of a research study on ED designed and conducted in the healthcare sector to determine the distress level in online diabetes community (patients, caregivers, support groups and doctors) as the targeted group. Diabetes is one of the largest global health concern with almost 451 million people (age 18-99 years) diagnosed with diabetes worldwide and this number is expected to increase to 693 million by 2045(Cho et al., 2017). Whereas, in Malaysia about 2.5 million people are diagnosed with diabetes in Malaysia. Studies shows that presence of Diabetes Distress (DD) is high in this community, arguing for a need to address DD thus, routine screening for depression and diabetes distress is essential for improving quality of life in diabetes community (Dieter & Lauerer, 2018). DD is also associated with feelings of Powerlessness, Hypoglycemia Distress, Negative Social Perceptions, Physician Distress and Family Distress.

Generally, people suffering from this chronic illness will have periodic contact with health professionals but they also need to have the skills, attitude, and support for self-management of their condition. Therefore, social networks as Facebook are an excellent resource for the patients since it helps to builds a bridge to connect different people who have a similar condition and similar experiences. It provides the environment and the tools for knowledge sharing and peer support. Patients' sharing information wrapped in their own sentiments and emotions, which is the driving force of sentiment analysis. This information could be beneficial to the healthcare centers or NGO's to identify the affected communities that are experiencing high distress level and help them to manage their distress level.

The focus of this paper is to study and implement Ensemble Method to enhance emotion detection for online diabetes community. This classification is then aimed to aid the detection of distress level in this diabetes community. This would be a major contribution to Healthcare Organizations and NGO's to help the affected communities that are identified to be distressed to avoid new outbreaks of depressions.



A dataset of 230000 Facebook post from a previous study was used for this study. This dataset goes through series of data cleaning where URL, posts less than 3 words, emoticons, urban slang (mamacita, papi, bling-bling) and non-English posts are removed. Total number of post after cleaning is 120000 and reactions are included. Emotion classification is carried out using Bagging (Ensemble method) and Support Vector Machine (Machine learning method) as classifiers. This classifier will classify the posts into 8 different emotion based on Plutchik's wheel of emotion which are joy, surprise, trust, disgust, fear, anger, sad and anticipation. Following this, distress level will be computed using the number of sad, angry, disgust and fear emotion detected. The final step is to evaluate the results using accuracy and precision metrics, human evaluation and benchmark studies.

REFERENCE

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. Pattern Recognition Letters, 93, 133-142. doi:https://doi.org/10.1016/j.patrec.2016.12.009

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. doi:https://doi.org/10.1016/j.asej.2014.04.011

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.), Mining Text Data (pp. 415-463). Boston, MA: Springer US.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice. doi:10.1016/j.diabres.2018.02.023

Dieter, T., & Lauerer, J. (2018). Depression or Diabetes Distress? Perspectives in Psychiatric Care, 54(1), 84-87. doi:doi:10.1111/ppc.12203

Breast Cancer Classification from Histopathology Images using Deep Neural Network

Ghulam Murtaza[1, 2*], Liyana Shuib[1*], Teh Ying Wah[1], Ghulam Mujtaba[1, 2], Ghulam Mujtaba[3]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
[2]Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan.
[3]PAF-KIET, City Campus 28-D, Block 6, P.E.C.H.S Karachi 75400, Sindh, Pakistan.

Corresponding Emails: gmurtaza@iba-suk.edu.pk, liyanashuib@um.edu.my, tehyw@um.edu.my, mujtaba@iba-suk.edu.pk, gmujtaba@mustaqim.com

Keywords: *Breast Cancer, Transfer Learning, Machine Learning, Hierarchical Classification, Histopathological Images, Image Classification, Convolutional Neural Network.*

BACKGROUND

Amongst the cancers, breast cancer has been reported the common types of cancer in women. Thus, several imaging technologies (including, X-rays, ultrasound, Magnetic Resonance Imaging, and histopathological images) are used to diagnose the breast cancer. In general, the histopathological images are mostly recommended for detailed analysis of breast cancer lesions. Thus, in recent years, a few researchers have proposed breast cancer classification models from histopathological images through the deep learning approaches. However, in those proposed models, there is still a need of improvement in the classification accuracy.

OBJECTIVE

This study aims to develop a binary classification model to classify the breast cancer into Benign or Malignant class from the histopathological images using deep neural network.

METHODS

To develop the proposed classification model, publicly available BreakHis dataset is used (Spanhol, Oliveira, Petitjean, & Heutte, 2016b). In addition, the dataset is divided into training-set and testing-set by ratio of 70% and 30% respectively. Furthermore, to overcome the issues of class imbalance and over-fitting of model training, image augmentation is employed on training-set only. Afterwards, the discriminative and powerful features are extracted from the images using pre-trained convolutional neural network and to form a master feature vector. The extracted master feature vector is then given as an input to six machine learning algorithms (namely, k-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, Linear Discriminant Classifier, and Linear Regression) to construct and evaluate the classification model. In addition, information gain feature reduction scheme is employed to obtain the subset of most discriminative and trivial features from master feature vector. Moreover, overall accuracy and area under the curve performance measures were used to evaluate the classification performance. Finally, the performance of proposed breast cancer classification model is compared to three state-of-the-art baseline breast cancer models (Nahid, Mehrabi, & Kong, 2018; Spanhol, Cavalin, Oliveira, Petitjean, & Heutte, 2017; Spanhol, Oliveira, Petitjean, & Heutte, 2016a) using the BreakHis dataset.

RESULTS

The experimental results showed that k-Nearest Neighbor outperformed other five machine learning algorithms by obtaining the highest classification accuracy of 95.48%. In addition, in kNN the optimal learning curve was observed by using 900 features out of 4096 features. Furthermore, the classification accuracy of proposed model is approximately 5%-10% better than the existing state-of-the-art baseline breast cancer models.

CONCLUSION

This study proposed deep neural network-based breast cancer classification model using histopathological images. The proposed model outperformed existing baseline models and can be deployed in real-time environment to serve as a second opinion for the pathologists. Moreover, it can exponentially reduce the diagnosis time and efforts taken by the pathologists in manually analyzing the histopathological images to classify the breast cancer. Finally, the approach of proposed model can be easily adapted in other related domains such as, lung cancer, prostate cancer, and bladder cancer. There are two major limitations of proposed model. First, it can only classify breast cancer into two classes namely, Benign and Malignant. Second, the model training time was extensive because it was trained on normal desktop machine having Corei7 processor and 8 GB RAM. Thus, in future, a classification model can be developed to classify breast cancer into more specific sub-classes of Benign (namely, Adenosis, Fibroadenoma, Tubular Adenoma, and Phyllodes Tumor) and Malignant (namely, Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma, and Papillary Carcinoma) classes using more powerful workstations having graphical processing units.

REFERENCES

Nahid, A.-A., Mehrabi, M. A., & Kong, Y. (2018). Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. *BioMed Research International, 2018*.

Spanhol, F. A., Cavalin, P. R., Oliveira, L. S., Petitjean, C., & Heutte, L. (2017). Deep Features for Breast Cancer Histopathological Image Classification.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016a). *Breast cancer histopathological image classification using convolutional neural networks.* Paper presented at the Neural Networks (IJCNN), 2016 International Joint Conference on.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *Ieee Transactions on Biomedical Engineering, 63*(7), 1455-1462.

Data Clustering using Ringed Seal Search

Shuxiang Zhang and Younes Saadi

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Email: younessaadi@gmail.com

The recent trends in data clustering approaches reveal that metaheuristics are very efficient due to their enormous capability to solve optimization problems with a fast convergence. In this talk, we discuss the capability of Ringed Seal Search (RSS) to solve data clustering. We also compare the performance of the RSS with three baseline algorithms: Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Cuckoo Search (CS), and finally we propose some important topics for further research. Clustering is considered as an unsupervised learning technique in which objects with a specific resemblance are classified into one group to form a cluster. However, the whole global convergence is considered to be slow (Nanda & Panda, 2014). Furthermore, it has been shown that clustering techniques based on metaheuristics are sensitive to the initial population settings, which can result in problems of local optima, as stated in (Nanda & Panda, 2014). Recently, a new metaheuristic approach called Ringed Seal Search (RSS) has been introduced. It is based on the behaviour of the ringed seal and its ability to find the best solution to escape predators (Saadi et al., 2016). The RSS requires fewer parameter settings and is characterized by an adaptive balance between exploitation and exploration, which makes the search able to escape local optimum traps easily. Moreover, the ability of RSS to solve clustering problems based on its fast convergence will be discussed. To verify the performance of the RSS algorithm, seven real benchmark datasets from the UCI Machine Learning Repository are used to compare the performance of RSS with other baseline algorithms such as GA, PSO, and CS. The RSS is referred to as a metaheuristic technique and was proposed by (Saadi et al., 2016) as an algorithm for searching for optimal solutions based on the best seal lair. The task of finding new solutions (new lairs) is particularly based on the sensitive search model, which is characterized by an adaptive balance between exploitation and exploration. As a result, the search task is divided into two states: normal and urgent. The seal's search is therefore designed to have two different patterns: normal search (normal state), where there is no noise, and urgent search (urgent state) in the case of noise. Every time a new good-quality lair is found, the seal will move into it. At the end, the lair with the best fitness value will be the term that RSS will optimize.

The clustering techniques group the objects into classes or clusters, which are formed based on a particular algorithm. The datasets that we consider contain numerical information on classes for each dataset. The RSS is proposed to build a new clustering approach, where the proposed approach is used to compute the optimal solution of the clustering objective function. The best lairs are considered as the best solution points, so the RSS search converges towards these points, finally forming the centres of the clusters. The proposed approach iterates until the stopping criterion is met.

The experimental results showed that the RSS-based data clustering shows an optimal performance compared to other clustering algorithms based on GA, PSO, and CS on most of the datasets. This performance is visualized in four figures and discussed, and the interpretation of the clustering results matched the convergence rate of the optimization algorithms used (Saadi et al., 2016). Furthermore, the results of the proposed approach demonstrate that using optimization algorithms featuring an optimal exploitation–exploration balance can lead to better clustering results.

In future research, we intend to evaluate other dissimilarity metrics related to clustering validation. Moreover, we plan to apply RSS-based data clustering on large datasets via MapReduce. This can give the RSS another dimension in solving clustering algorithms via big data platforms.

REFERENCES

Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation, 16*, 1-18.

Saadi, Y., Yanto, I. T. R., Herawan, T., Balakrishnan, V., Chiroma, H., & Risnumawan, A. (2016). Ringed Seal Search for Global Optimization via a Sensitive Search Model. *PloS one, 11*(1), e0144371.

A Conceptual Model of Foreign Students Profiling

R. Renugah[1], Suraya Hamid[1] and Abdullah Gani[2]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
[2]Department of Computer Systems and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Emails: renugah@siswa.um.edu.my, suraya_hamid@um.edu.my, abdullah@um.edu.my

Keywords: *foreign students, student profiling, social network analysis, data analytics, theory of planned behaviour, social cognitive theory*

INTRODUCTION

The rise of the foreign student's enrolment is estimated at 7.2 million in the year 2025 from 1.8 million in the year 2000 (Bohm, Davis, Meares, & Pearce, 2002). Malaysia is striving towards establishing itself as a regional education hub by 2020 through research and development programme was outlined in the 9th Malaysian Plan (2006-2010) (Cheng, Mahmood, & Yeap, 2013; Knight & Morshidi, 2011). It all began when Ministry of Higher Education (MOHE) launched of The National Higher Education Strategic Plan 2020 (NHESP) in August 2007 with the notion of reforming the nation's higher education. The fifth initiative-intensifying internationalization from the Phase 2 (2011-2015) looking forward in developing Malaysia as an education hub in the region. In addition, MOHE also aiming to recruit 200,000 foreign students and to achieve one of the top six countries choices for the foreign students by the year 2020 (Lewis, 2016; Shahijan, Rezaei, & Amin, 2016).

In mid-1997, the Asian financial crisis has triggered economic chaos and deteriorate the economy but it was a contradictory implication in education especially higher education (Asari, Muhamad, & Khalid, 2017) as it encouraged potential foreign students to pursue higher education in Malaysia due to the affordable fees compared to other countries like USA and UK (Tan, 2002). Middle East country student's encountered difficulties in pursuing the education in the USA after the 9/11 attack forcing the students to seek more practical and affordable options (Abd Aziz & Abdullah, 2014; Sirat, 2008). Besides, Malaysia is known for its diverse culture despite being an Islamic country, also has attracted students not only from Arab, Africa but also from Asia.

Education Malaysia Global Services (EGMS) was launched in 2013 to streamline as well as to reduce the bureaucracy dilemma as every foreign student applications will go through the academic screening process before being forwarded to the Immigration Department for the Visa Approval Letter (VAL) (Keong, Naim, & Zamri, 2014; StudyMalaysia.com, 2017). It is stated that the EGMS establishment could minimize the last minute approvals which affected the foreign student admissions by 15% (Keong et al., 2014). Besides, it is also driven by the establishment of foreign universities and private college's branch campuses in Malaysia (Abd Aziz & Abdullah, 2014; Knight & Morshidi, 2011; Tham, 2013). Malaysia has attracted the foreign students as their top choices due to many aspects such as its affordable tuition fees, usage of English as instruction medium, lower cost of living and also home to multi-ethnic diversity (Manjula & Slethaug, 2011).

There have been some hitches among the foreign students in Malaysia such as visa and drug misuse and anti-social activities despite going through a stringent process before approval being issued to these foreign students (Umar, Noon, & Abdullahi, 2014). Some of these dubious foreign students are not listed in the suspect record, therefore making it impossible to identify and distinguish them as genuine foreign students or not. The rising concern over terrorism may taint the nation's image in the eye of the world, therefore the authorities ought to take a pro-active action to overcome the student visa abuse by the dubious party. The current procedure is still lacking and failed to distinguish these dubious foreign students due to few factors as the information is merely provided by the students; and it only involves a one time

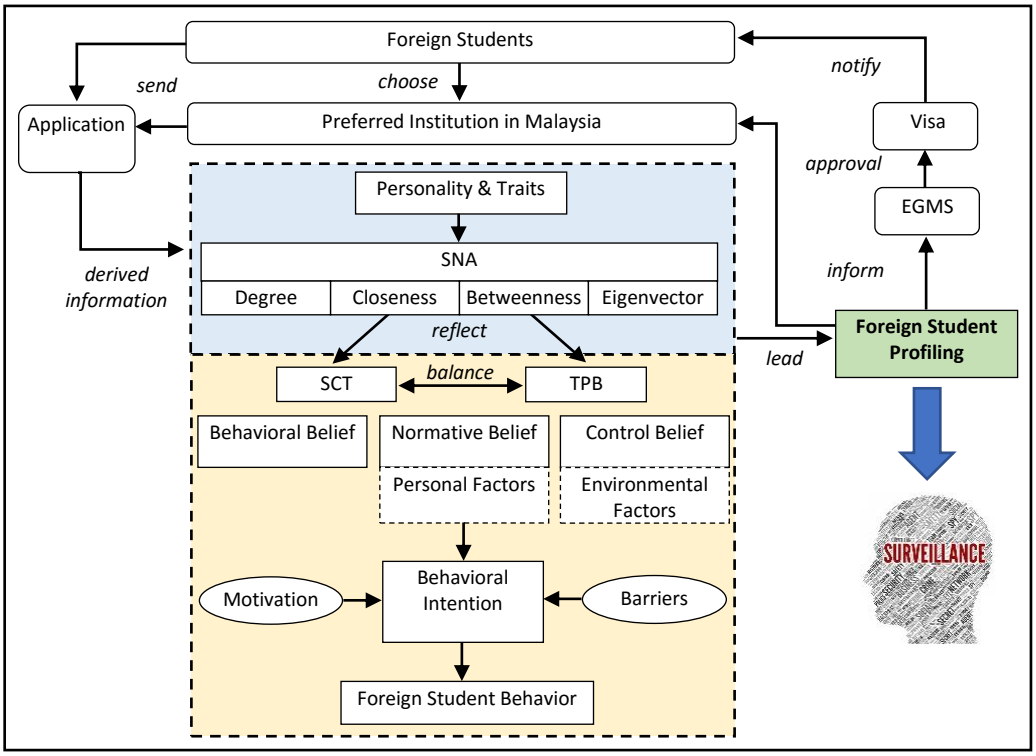
screening process – during the application stage; some of these students are not recorded in the suspect list; also the human behaviours may change at any time under some circumstances; besides there is no continuation scrutiny involved after the admission at the institution.

Therefore, this study is intended to seek an answer on how to identify and distinguish the potential foreign student's right from the beginning of their application. Our main aim is to propose a method; profiling the foreign students by exploiting the richness data available in social media as it may provide us some notion in a different point of view. Some of the online content can be found to be offensive, notorious or confrontational and also could map the associations with the professional or confrontational groups. Being said that, profiling could be a guidance by connecting the behaviour pattern to the individual characteristics.

MODEL DEVELOPMENT

The foreign students' features will serve as a basis for the conceptual model of the foreign student profiling. The features extracted will include their personal traits, demographic information and also the social aspects. Then, these components will be applied to social network analysis (SNA) using the main four centrality measures. Then, theory of planned behaviour (TPB) and social cognitive theory (SCT) elements such as i) normative belief and subjective norms - reflects on the individual's perception in performing a behaviour; ii) control beliefs and perceived behavioural control – reflects on the factors that facilitate or hinders in performing a behaviour and also about the individual's level of difficulty in performing a behaviour; iii) behavioural intention and behaviour reflects on the individual's readiness of performing a behaviour which based on the attitude, subjective norm and perceived behavioural control, as well as the individual's reaction towards the behaviour will be applied. The TPB and SCT will be mapped together based on the moderating aspects; motivation and barriers to determine the foreign students behavioural in addition to their personal traits.

In a nutshell, the essential information from the foreign student's application will be extracted before the SNA, SCT and TPB applied to disseminate their unique profile. From here, the authorities could decide the approval of these foreign student's application as well as their visa application. The proposed conceptual model of foreign student profiling is depicted in Figure 1 below.



SIGNIFICANCE

This study is aimed to overcome these problems by profiling these foreign students by using the richness of data available in social media. The proposed profiling model could be an added value along with the current conventional approaches. Apart from that, it is expected to help the relevant authorities to monitor and conduct the scrutiny periodically or at any point, if any issues raised involving the foreign students in the country. Moreover, it is also could assist the authority to improve their services by making real-time decision-making.

ACKNOWLEDGEMENT

The main author acknowledges Public Service Department of Malaysia (JPA) for the Hadiah Latihan Persekutuan (HLP) scholarship to undertake this research as part of Ph.D. work.

REFERENCES

Abd Aziz, M. I., & Abdullah, D. (2014). Finding the next 'wave' in internationalisation of higher education: focus on Malaysia. *Asia Pacific Education Review, 15*(3), 493-502. doi:10.1007/s12564-014-9336-7

Asari, F. F. A. H., Muhamad, S., & Khalid, P. Z. M. (2017). Globalisation and Liberalisation of Malaysian Higher Education. *ESTEEM Journal of Social Sciences and Humanities, Volume 1*, pp. 1-14.

Bohm, A., Davis, D., Meares, D., & Pearce, D. (2002). Global student mobility 2025: Forecasts of the global demand for international higher education. *IDP Education Australia*.

Cheng, M. Y., Mahmood, A., & Yeap, P. F. (2013). Malaysia as a regional education hub: a demand-side analysis. *Journal of Higher Education Policy and Management, 35*(5), 523-536.

Keong, Y. C., Naim, S., & Zamri, N. D. M. (2014). Online news report headlines of education Malaysia global services. *Jurnal Komunikasi, Malaysian Journal of Communication, 30*(2).

Knight, J., & Morshidi, S. (2011). The complexities and challenges of regional education hubs: Focus on Malaysia. *Higher Education, 62*(5), 593.

Lewis, V. (2016). Embedding marketing in international campus development: lessons from UK universities. *Perspectives: Policy and Practice in Higher Education, 20*(2-3), 59-66.

Manjula, J., & Slethaug, G. (2011). *The business of education: Improving international student learning experiences in Malaysia.* Paper presented at the 14th International Business Research Conference, Dubai UAE.

Shahijan, M. K., Rezaei, S., & Amin, M. (2016). International students' course satisfaction and continuance behavioral intention in higher education setting: an empirical assessment in Malaysia. *Asia Pacific Education Review, 17*(1), 41-62.

Sirat, M. (2008). The Impact of September 11 on International Student Flow Into Malaysia: Lessons Learned. *International Journal of Asia-Pacific Studies, 4*(1).

StudyMalaysia.com. (2017). Becoming An International Student In Malaysia And Immigration Procedures. Retrieved from https://studymalaysia.com/international/applying-entering-malaysia-to-study/becoming-an-international-student-in-malaysia

Tan, A. M. (2002). *Malaysian private higher education: Globalisation, privatisation, transformation and marketplaces*: Asean Academic PressLtd.

Tham, S. Y. (2013). Internationalizing Higher Education in Malaysia:Government Policies and University's Response. *Journal of Studies in International Education, 17*(5), 648-662. doi:10.1177/1028315313476954

Umar, A., Noon, N. A. M., & Abdullahi, M. (2014). Challenges confronting African students in Malaysia: A case of postgraduate Nigerian students at International Islamic University Malaysia (IIUM) Kuala Lumpur. *Journal of African Studies and Development, 6*(9), 161-168.

Big Data in Urban Planning

Rosilawati Zainol
Centre for Sustainable Urban Planning and Real Estate (SUPRE), Faculty of Built Environment, University of Malaya
Centre for Civilisational Dialogue, University of Malaya

Corresponding Email: rosilawatizai@um.edu.my

INTRODUCTION

Big data has been around since 1950's. However, its usage is very limited compared to today's standard. Recently, it has gained more attention. Its benefits to the urban studies and planning practices have been realized by urban planners.

PROBLEM STATEMENT

Urban planners always resort to the traditional method of data collection. In employing a quantitative approach in any study, their main sources of data are through field work, observation, questionnaire survey and census data. According to Batty (2013), census data is based on the enumeration of the population. This data is collected every ten years which misses a lot of necessary action. Similarly, fieldwork and observation cover a limited area. Likewise, a questionnaire survey is merely perceptions by respondents. On the other hand, big data obtained from social media, public transportation concession cards and mobile phones can yield real data which can relay routine sensing and reveal enormous heterogeneity data in cities. Batty (2016) defines big data in the context of urban planning in a 3Vs model: volume, velocity and variety. He further adds two more Vs: variability and veracity. The availability of these data will enable urban planners to make better planning decisions. However, many urban planners are unaware of the importance of big data. Thus, this review intends to highlight the importance of using big data in urban planning decision making to produce a more sustainable city policy.

METHODOLOGY

This study employs review of the literature as its main method. Its aim is to highlight the summary of outputs of previous studies. The term 'big data' and 'urban planning' are the two keywords used to search literature related to this study. Google scholar database was used to search for the literature. Since big data deals with the internet of things, only literature published since 2013 are included in this study. Only 98 articles are found to be related to this study according to the keywords through search done in Google Scholar. However, only 26 articles are from journal articles. In this study of journal articles are included in the review.

FINDINGS

Findings show scholars have identified the importance of big data in urban planning and governance. Emphases are placed more on handling urban issues such as low carbon concept in block planning; mobility space distribution, shifting from medium and long-term planning to short term planning; greenspaces; determining best location for power line and substation; urban governance; diurnal commuting flows; detecting hotspots; framework for data use for smart cities; awareness & property prices, advertisement effect; urban heat island in planning practice; behaviour data acquisition and analysis, spatial analysis, plan making and management application and new methodologies; trip planning; subway station functions; human mobility patterns; change urban population at different levels; noise management; validation and many others. Besides that, some scholars identified the sources of big data such as mobile phone, crowdsourcing, smart cards, social media platform, residents' complaints, open data platforms and acoustic sensor networks. Some scholars even created some applications that can gather and analyse big data.

IMPLICATIONS

Acquisition of big data by urban planners is crucial. They need to grasp this skill to be at par with other disciplines in capturing and analysing big data for their short-term decision making and policy making. Today's world requires urban planners to act quickly and efficiently. This study is a preliminary study on creating awareness on the usage and importance of big data for urban planners.

REFERENCES

Arafah, Y., & Winarso, H. (2017). Redefining smart city concept with resilience approach. *IOP Conference Series: Earth and Environmental Science, 70*(1), 012065.

Batty, M. (2017). Data about cities: Redefining big, recasting small. In T. P. L. Rob Kitchin, Gavin McArdle (Ed.), *Data and the City* (pp. 13). London: Routledge

Batty, M. (2016). Big data and the city. *Built Environment, 42*(3), 321-337.

Coletta, C., Heaphy, L., Perng, S.-Y., & Waller, L. (2017). Data-driven Cities? Digital Urbanism and its Proxies: Introduction. *Tecnoscienza Italian Journal of Science & Technologies Studies, 8*(2), 5-18.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1), 2053951716631130. doi:10.1177/2053951716631130

Kong, X., Li, M., Li, J. et al. (2018). *CoPFun: an urban co-occurrence pattern mining scheme based on regional function discovery.* World Wide Web https://doi.org/10.1007/s11280-018-0578-x

Leao, S., Lieske, S., Conrow, L., Doig, J., Mann, V., & Pettit, C. (2017). Building a National-Longitudinal Geospatial Bicycling Data Collection from Crowdsourcing. *Urban Science*, *1*(3), 23. MDPI AG. Retrieved from http://dx.doi.org/10.3390/urbansci1030023

Mao, D., Li, Z., Li, H., & Wang, F. (2018, 15-17 Jan. 2018). *Bike-Sharing Dynamic Scheduling Model Based on Spatio-Temporal Graph.* Paper presented at the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp).

Navarro, J. M., Tomas-Gabarron, J. B., & Escolano, J. (2017). A Big Data Framework for Urban Noise Analysis and Management in Smart Cities. *Acta Acustica united with Acustica*, *103*(4), 552-560. doi:10.3813/AAA.919084

Selva Royo, J. R., Mardones, N., & Cendoya, A. (2017). *Cartographying the real metropolis: A proposal for a data-based planning beyond the administrative boundaries*.

Song, Y., Huang, B., Cai, J., & Chen, B. (2018). Dynamic assessments of population exposure to urban greenspace using multi-source big data. *Science of The Total Environment*, *634*, 1315-1325. doi:https://doi.org/10.1016/j.scitotenv.2018.04.061

Zhu, Y. (2018) *Estimating the activity types of transit travelers using smart card transaction data: a case study of Singapore.* Transportation. https://doi.org/10.1007/s11116-018-9881-8

Developing Student Engagement Model Using Learning Analytics

Shahrul Nizam Ismail and Suraya Hamid

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Emails: shahrulnizam.ismai@gmail.com, suraya_hamid@um.edu.my

Keywords: *Learning Analytics, Student Engagement Model, Learning Management System, Online Learning Analysis*

INTRODUCTION

Learning Analytics is the measurement, collection and analysis of learners' related-data for understanding and learning optimization purposes. Trace data and log data that are left behind in the Learning Management System (LMS) or virtual learning environment (VLE) are the main sources of data for the analytics purpose (JIl-Hyun et al. 2015). The analytics will produce an insight into the learning log data to provide better decision making process to the various stakeholders. The log data and student behavior are the relationships that can be identified using learning analytics (Rienties & Toetenel, 2016). Learning Analytics also have been used in predicting student's performance, dropout rate, final mark and grade, retention of the students in LMS, and early warning system for the students throughout their study in university (Casey & Azcona, 2017). The implementation of Learning Analytics helps university administration and faculty members to monitor the progress of the students and know the rate of their progress in the process of learning.

PROBLEM STATEMENT

Every data produced from the Learning Management System is valuable and researchers must know the category of data that can provide contribution to their research. The engagement between lecturers and students is important to develop a healthy relationship not only in class but also in the virtual learning (Ma et al., 2015). The response from the students in an activity can discover their eagerness towards virtual learning and increase the activity lesson from the classroom. Students acceptance toward the Learning Management System (Jiseong et al., 2016) or virtual learning environment is positive as they can easily capture the role of the system and usefulness of the Learning Management System, but there are many factors that encourage the lack of engagement in the Learning Management System (Holmes, 2018; Zhang et al.,2017). Therefore this research tends to develop an engagement model of the students in LMS which in return can enrich the usage of LMS.

RESEARCH DESIGN

The dataset are collected from a Public University in Malaysia through its Center of Information Technology and the log data will be gotten from one of the faculties in the University. The log data comprise of data gotten from the year 2014 up to 2016. We will analyze the log activity from the Learning Management System. The activity we shall analyze shall comprise of Forum, Discussion, Assignment Submission, Resource download, and Time spent in the system. The correlation analysis and multiple regression analysis shall be applied to the variables to determine the significance of the activity by ranking the most engaged activity. The analysis will be conducted using the SAS Enterprise Miner 9.4. The model developed will be validated via an interview with selected lecturers and faculty administration staff. The analysis purpose is to determine the engagement in terms of Teamwork Engagement (group assignment, presentation), Students Engagement (access the material, number of downloads, number of logins), and Engagement with the lecturers (discussion, assignment, tutorial, and quiz).

FINDINGS

From the initial study of the literature review, the activity on the Learning Management System that have a significance on the determination of the student's interaction and engagement are listed in the table below. They consist of elements of engagement in LMS, existing student engagement model and learning analytics implementation in higher education institutions.

Table 0.2: Engagement Elements in LMS

Elements	Descriptions	Example
Time Spent	Number of login, total time in the LMS	(Firat, 2016; Abzug, 2015)
Assignment	Submission of the assignment within the due date	(Strang, 2016; Nguyen, 2017)
Discussion	Active participation (one-to-one, one-to-many, many-to-many)	(Chen et al., 2018; Holmes, 2018)
Resource	Number of downloads and access to the resources	(Rhode, 2015; Aljarrah et al., 2018)

Existing Student Engagement Model includes: a) Ballard & Butler (2016) Conceptual Model: focus on Engagement Profile. b) Zhang et al. (2017) Research Model: Teamwork Engagement. c) Ma et al. (2015) Teaching and learning interaction activity model. Learning Analytics in Higher Education in other countries include; a) the predictive model identifies students who may require support, by Edith Cowan University, Australia. b) Using total VLE hits is an excellent starting point for predictive modelling and early warning systems by California State University, USA. c) Predictive analytics are being developed to model student progression at individual module levels and institutional levels by Open University, United Kingdom.

The initial model will be developed based on the synthesis of the above findings and also the analytics insight from the log data will be used to develop the Student Engagement Model.

IMPLICATIONS

The model that will be developed can produce an impact factor to increase the interaction and engagement between lecturers and students in LMS as well as to achieve the Malaysia Blueprint Education on Globalized E-Learning by increasing the tendency used on LMS sites.

ACKNOWLEDGEMENT

This research is financially supported by University of Malaya, Bantuan Khas Penyelidikan (BKP), under research grant BKS083-2017.

REFERENCES

AlJarrah, A., Thomas, M. K., & Shehab, M. (2018). Investigating temporal access in a flipped classroom: procrastination persists. International Journal of Educational Technology in Higher Education, 15(1). https://doi.org/10.1186/s41239-017-0083-9

Ballard, J., & Butler, P. I. (2016). Learner enhanced technology: Can activity analytics support understanding engagement a measurable process? Journal of Applied Research in Higher Education, 8(1), 18–43. Retrieved from https://doi.org/10.1108/JARHE-09-2014-0074

Casey, K., & Azcona, D. (2017). Utilizing student activity patterns to predict performance. International Journal of Educational Technology in Higher Education, 14(4). https://doi.org/10.1186/s41239-017-0044-3

Chen, B., Chang, Y. H., Ouyang, F., & Zhou, W. (2018). Fostering student engagement in online discussion through social learning analytics. *Internet and Higher Education*, *37*, 21–30. https://doi.org/10.1016/j.iheduc.2017.12.002

Firat, M. (2016). Determining the effects of LMS learning behaviors on academic achievement in a learning analytic perspective. Journal of Information Technology Education Journal of Information Technology Education: Research, 15(15), 75–87. https://doi.org/10.28945/3405

Holmes, N. (2018). Engaging with assessment: Increasing student engagement through continuous assessment. *Active Learning in Higher Education*, *19*(1), 23–34. https://doi.org/10.1177/1469787417723230

Ma, J., Han, X., Yang, J., & Cheng, J. (2014). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. Internet and Higher Education, 24, 26–34. https://doi.org/10.1016/j.iheduc.2014.09.005

Nguyen, V. A. (2017). The Impact of Online Learning Activities on Student Learning Outcome in Blended Learning Course. *Journal of Information & Knowledge Management*, 1750040. https://doi.org/10.1142/S021964921750040X

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behavior, satisfaction and performance: A cross-institutional comparison across 151 modules. Computers in Human Behavior, 60, 333–341. https://doi.org/10.1016/j.chb.2016.02.074

Sclater, N., Mullan, J. & Peasgood, A., 2016, Learning Analytics in Higher Education: A Review of UK and International Practice, Jisc. jisc.ac.uk/reports/learning-analytics-in-higher education

Person Abnormal Behaviour Identification through Posting Images in Social Media

Divya Krishnani[1], Palaiahnakote Shivakumara[2], Tong Lu[3] and Umapada Pal[4]
[1]International Institute of Information Technology, Naya Raipur, Chhattisgarh, India.
[2]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[3]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China.
4Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.

Corresponding Emails: divya16100@iiitnr.edu.in, shiva@um.edu.my, lutong@nju.edu.cn, umapada@isical.ac.in

Keywords: *Social media, Face detection and recognition, Emotions detection, Abnormal behaviour identification.*

As day passes, use of social media increases with high speed and it has becoming an integral part of person life because it plays a vital role in many fields, such as Education, Industry, Agriculture, Business, Communication etc. (Injadat, Salo, & Nassif, 2016; Mabrouk & Zagrouba, 2017). As a result, one can expect unexpected collection of data and usage every day, which results in big data. Therefore, big data requires data analytics or data scientist to extract useful information according to requirement and applications from diversified huge data. One such attempt is to identify the person behaviour based on what he/she is posting images in social media(Hsu, Chuang, Huang, Teng, & Lin, 2018; Liu, Preotiuc-Pietro, Samani, Moghaddam, & Ungar, 2016). It is a fact that posting images provide better information compared to edited text by the persons. In addition, images can easily communicate emotions, feelings, intensions, bonding etc. as it does not require any specific language knowledge unlike edited text that requires language knowledge. Therefore, persons use social media for expressing their own feelings, emotions, behaviour by posting respective images (Injadat, Salo, & Nassif, 2016; Mabrouk & Zagrouba, 2017). At the same time, the same way can be used for sending message, which can be joyful, threatening, daring, scaring, attracting to particular person/group/public(Tiwari, Hanmandlu, & Vasikarla, 2015). This situation demands to develop a tool, which can identify the abnormal and normal behaviour by analysing the images posted by different persons automatically such that one can prevent evil impacts on society. Hence, we propose to develop a model for identifying person behaviour through images posted in social media.

There are methods and models available in literature for identifying the normal and abnormal behaviour based on analysing edited text in the field of natural language processing and data mining (Injadat, Salo, & Nassif, 2016; Mabrouk & Zagrouba, 2017). However, these methods work well for text based images but not the images posted in social media. Similarly, we can also find several methods for normal, abnormal behaviour, emotions and expressions identification using face recognition and facial features (Tiwari, Hanmandlu, & Vasikarla, 2015). However, most methods work well when we give individual faces. In addition, the methods are useful for the images captured by high-resolution camera but not the images posted in social media where one can expect poor quality and low-resolution images. Further, the methods can identify successfully emotions, expressions but not action in the image, such as Bullying, Threatening, Depression, Sarcasm and Psychopath. Hence, there is a need for developing a new model for the images posted in social media.

In this work, we consider six classes, namely, Normal (Extroversion) which generally contains smile faces and general information, Bullying, which contains sad, angry face along with action information, Threatening, which contains scary and anger faces, Neuroticism-Sarcastic which contains face with unique smile, Neuroticism-Depression which contains sad face with different action and Psychopath which contains face with unnatural information. We propose a new model based on Hanman Transform (Grover & Hanmandlu, 2018), which combines facial features and background information to identify the above-mentioned six behaviours in this work. It is noted that in all the classes, the content of the face is

prominent information, which gives clue for the different behaviour and the context of foreground-background of the images, which gives clues for action in the images. For face region detection, we use existing well-known Multi-task Cascaded Convolutional Networks (MTCNN) framework (Zhang, Zhang, Li, & Qiao, 2016) and for the components in the background, we use Canny edge image of the input image. It is true that the above observations reflect in the face region and other components. Therefore, to extract such changes in the face region and components, we introduce Hanman Transform, which is good for studying uniform where there are no ambiguities and non-uniform where there are uncertainties. Based on Hanman values, the proposed model separates foreground (which represent high Hanman values) and background components (which represent low Hanman values). In order to extract the context features, the proposed model finds the relationship between high and low Hanman values of respective components through fusion criterion (Xu, Wang, & Chen, 2016). This process results in feature matrix. The feature matrix is subjected to the classifier for final classification. The proposed model is evaluated by conducting experiments on large data in terms of recall, precision and f-measure and classification rate. The effectiveness and usefulness of the proposed model is demonstrated by comparing with the existing methods.

REFERENCES

Grover, J., & Hanmandlu, M. (2018). The fusion of multispectral palmprints using the information set based features and classifier. *Engineering Applications of Artificial Intelligence, 67*, 111-125.

Hsu, S.-C., Chuang, C.-H., Huang, C.-L., Teng, R., & Lin, M.-J. (2018). *A video-based abnormal human behavior detection for psychiatric patient monitoring.* Paper presented at the Advanced Image Technology (IWAIT), 2018 International Workshop on.

Injadat, M., Salo, F., & Nassif, A. B. (2016). Data mining techniques in social media: A survey. *Neurocomputing, 214*, 654-670.

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. H. (2016). *Analyzing Personality through Social Media Profile Picture Choice.* Paper presented at the ICWSM.

Mabrouk, A. B., & Zagrouba, E. (2017). Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*.

Tiwari, C., Hanmandlu, M., & Vasikarla, S. (2015). *Suspicious Face Detection based on Eye and other facial features movement monitoring.* Paper presented at the Applied Imagery Pattern Recognition Workshop (AIPR), 2015 IEEE.

Xu, X., Wang, Y., & Chen, S. (2016). Medical image fusion using discrete fractional wavelet transform. *Biomedical Signal Processing and Control, 27*, 103-111.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503.

Modification of an Encryption Scheme Using the Laplace Transform

Roberto Briones
Heriott Watt University, Malaysia

Corresponding Email: r.briones@hw.ac.uk

Keywords: *Laplace transform, Maclaurin expansion, plaintext, cyphertext, security key, modular arithmetic*

Hiwarekar (2012) recently introduced a new scheme in cryptography whose construction is based on the Laplace transform. The encryption process is based on pre-selecting an underlying C^{∞} function $f(rt)$, writing out its Maclaurin series, multiplying said series with t^k, multiplying term-wise the numerical codes of the letters of the plaintext with the coefficients of the first terms of the resulting series, and finally determining the Laplace transform of the subsequent finite series, with a view of utilizing the initial coefficients of the last series as the basis of the cyphertext. (Gupta and Mishra, 2014) posited that the single-iteration procedure offers a weak encryption scheme by showing that cyphertext messages can be decrypted by elementary modular arithmetic, and stating that the procedure is independent of the Laplace transform. This paper discusses a way of strengthening the purported source of weakness of the Hiwarekar scheme, and modifies the initial step of the encryption process, consequently giving rise to two passwords for a single iteration procedure. In the end, the receiver receives from the sender a security key, and a second one called a subscript key, which is based on random selection of coefficients from the generated series of Laplace transforms. The additional key strengthens the security of the cyphertext generated by the algorithm.

REFERENCES

Gençoğlu, M.T. (2017). Cryptanalysis of a New Method of Cryptography using Laplace Transform Hyperbolic Functions. Communications in Mathematics and Applications **8** (2), 183–189.

Gupta, P., Mishra, P.R. (2014). Cryptanalysis of "A New Method of Cryptography Using Laplace Transform". In: Pant, M., Deep, K., Nagar, A., Bansal, J., (eds) Proceedings of the Third International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing **258**, 539 – 546. Springer, New Delhi.

Hiwarekar, A.P. (2012). A new method of cryptography using Laplace Transform. International Journal of Mathematical Archive **3** (3), 1193 – 1197.

Hiwarekar, A.P. (2015). Application of Laplace Transform for Cryptography. International Journal of Engineering & Science Research **5** (4), 129 – 135.

Implicit Feedback and Comparative Analysis of Online IR Evaluation Methods

Sinyinda Muwanei[1],Sri Devi Ravana[1],Hoo Wai Lam[1],Douglas Kunda[2]
[1]*Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia*
[2]*Department of Computer Science, Mulungushi University*

Corresponding Emails: smuwanei@gmail.com, sdevi@um.edu.my

Keywords: *Online information retrieval, implicit feedback, evaluation*

INTRODUCTION

Implicit feedback plays a critical role in inferring relevance to documents in the field of information retrieval. Retrieval tasks that employ implicit signals include but not limited to ranking, evaluation, automatic tuning of retrieval functions, document expansion and personalized search. Implicit signals investigated in literature include click through data, dwell time and eye tracking. Few works however exist on the study of how reliable these implicit signals are in various retrieval tasks outlined above.

According to Joachims et al (2005) even though clicks are biased, they are more informative than manual relevance judgements and user preferences derived from them are accurate on average.

According to Shinoda, (1994) the reading time is indicative of interest of users when reading news stories. In agreement to this conclusion is Claypool et al (2001) who conclude that reading time as well as the amount of scrolling can predict relevance in web browsing, while individual mouse movement and clicks were ineffective in predicting relevance. Similarly, Fox et al (2005) show in their study that the overall time a user interacts with a search engine, as well as the number of clicks, are indicative of user satisfaction with the search engine. Not only can dwell time and click through data infer relevancy but eye movements. Salojärvi et al (2003) used measures of pupil dilation to infer the relevance of online abstracts, and found that pupil dilation increased when fixated on relevant abstracts.

One information retrieval application area for implicit signals is evaluation of information retrieval systems. In this paper an experiment is conducted to compare the effectiveness of some multileaving information retrieval ranker evaluation methods. These are Team-Draft Multileaving (TDM) introduced by Schuth et al. (Schuth, Sietsma, Whiteson, Lefortier, & de Rijke, 2014), Sample-Scored-Only Multileaving (SOSM) introduced by Brost et al. (Brost, Cox, Seldin, & Lioma, 2016), Probabilistic Multileaving (PM) introduced by Schuth et al (Schuth et al., 2015) and Pairwise Preference Multileaving introduced by Oosterhuis et al. (Oosterhuis & de Rijke, 2017)

EXPERIMENTAL SETUP AND RESULTS

The OHSUMED dataset is used and is based on the query log of the search engine on the MedLine abstract database and contains 106 queries. User interactions are vital for online IR evaluation experiments hence the cascade model presented by Guo et al.(Guo, Liu, & Wang, 2009) is used to model user behaviour. Implementation for Pairwise Preference Multileaving introduced by Oosterhuis & de Rijke (Oosterhuis & de Rijke, 2017) paper is utilised for the experiment. However, for the experiment in this paper seven rankers and one thousand impressions are set. Results are shown diagrammatically below under informational, navigational and perfect click models and they show that the bin error rates on average for multileaving methods tend to reduce as the number of impressions increase for multileaving methods to as low as below 15% especially for pair wise multileaving method. This is quite impressive even though only clicks were used to model user behaviour.

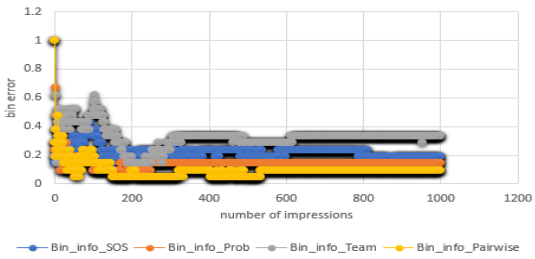


Figure 1 – Multileaving Evaluation Methods bin error comparison under the informational click model

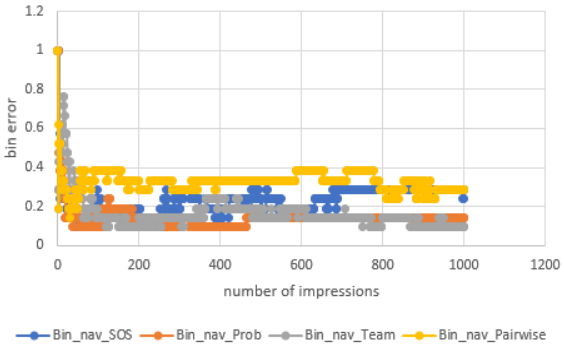


Figure 2 – Multileaving Evaluation Methods bin error comparison under the navigational click model

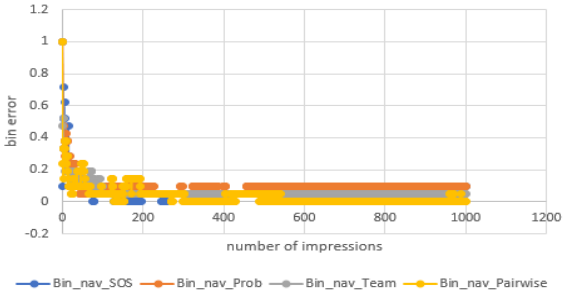


Figure 3 – Multileaving Evaluation Methods bin error comparison under the perfect click model

CONCLUSION

Online IR is still in its infancy and more effort is required from the research community to bring this field to maturity. From the experiments above it is evident that online IR evaluation methods only consider click implicit signals on documents as the only inference of document relevancy and their binary error rates for the selected data set are quite impressive. Review of implicit signals did show that inferency of document relevancy should better be done with combination of implicit signals. Therefore in future information retrieval researchers, should consider developing IR evaluation methods that have even lower binary error rates while incorporating dwell time, eye tracking and facial expressions in inferences of document relevancy. Consequently this might lead to solutions of some open questions still in the multi duelling bandits problem when applied to IR ranker evaluations.

REFERENCES

Brost, B., Cox, I. J., Seldin, Y., & Lioma, C. (2016). An Improved Multileaving Algorithm for Online Ranker Evaluation. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '16*, *3*(1), 745–748. https://doi.org/10.1145/2911451.2914706

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. *Proceedings of the 6th International Conference on Intelligent User Interfaces - IUI '01*, 33–40. https://doi.org/10.1145/359784.359836

Fox, S., Karnawat, K., Mydland, M., & Dumais, S. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, *23*(2), 147–168. https://doi.org/10.1145/1059981.1059982

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '05*, *51*(1), 154. https://doi.org/10.1145/1076034.1076063

Oosterhuis, H., & de Rijke, M. (2017). Sensitive and Scalable Online Evaluation with Theoretical Guarantees. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, 77–86. https://doi.org/10.1145/3132847.3132895

Salojärvi, J., Kojo, I., Simola, J., & Kaski, S. (2003). Can relevance be inferred from eye movements in information retrieval. *Proceedings of the 4th Workshop on Self-Organizing Maps (WSOM 2003)*, (September), 261–266. https://doi.org/10.1145/1460096.1460120

Schuth, A., Bruintjes, R.-J. R., Büttner, F., van Doorn, J., Groenland, C., Oosterhuis, H., … de Rijke, M. (2015). Probabilistic Multileave for Online Retrieval Evaluation. *Sigir '15*, (1), 2–5. https://doi.org/10.1145/2766462.2767838

Schuth, A., Sietsma, F., Whiteson, S., Lefortier, D., & de Rijke, M. (2014). Multileaved Comparisons for Fast Online Evaluation. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, 71–80. https://doi.org/10.1145/2661829.2661952

Shinoda, Y. (1994). Sigir '94, (August 1994). https://doi.org/10.1007/978-1-4471-2099-5

Analyzing and visualizing Thoracic Surgery Data Set

Samar Bashath[1] and Amelia Ritahani Ismail[2]

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

Corresponding Emails: samarsalim7076@gmail.com, amelia@iium.edu.my

INTRODUCTION

In recent year, the difficulty for creating model increasing explosively as the data to be analyzed growing. When the attempt is classifying or creating a model for a specified problem, the various technique could be used, but dealing with problems contain a high number of features is a very challengeable task (Mirkin, 2011). In addition, the overfitting problem could affect the accuracy of model prediction (Bilbao & Bilbao, 2018). The proper way is highly required to prevent the effects of overfitting.

THE AIM OF STUDY

This paper focuses on the analyzing and selecting the data using statistical methods before creating the model.

METHODOLOGY

We chose the real data which **Thoracic Surgery Data Data from the UCI- machine learning resporsiity** (Zięba, Tomczak, Lubicz, & Świątek, 2014) for studying, analyzing and visualizing the data**.** The data is for determining whether the patient could live for one year after lung cancer operation. The following steps were conducted.

Step 1-Initial data exploration

There are three scale variables: Age, volume, capacity. In addition, the `dataSet` observations include nominal variables: Diagnosis-specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1). Performance_status - Zubrod scale ((PRZ2,PRZ1,PRZ0)- Pain_before_surgery: (T,F) - Haemoptysis_before_surgery (T,F) - Dyspnoea_before_surgery (T,F) - Cough_before_surgery (T,F) -Weakness_before_surgery (T,F) , Size_of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13) - Type _2 DM - diabetes mellitus (T,F) -MI up to 6 months (T,F) -PAD - peripheral arterial diseases (T,F) -Smoking (T,F) - Asthma (T,F) -one_year_status (T)rue value if died (T,F).

Step 2- Correlation between scales variables and one-year status

The ANOVA test (Bewick, Cheek, & Ball, 2004)() was used to determine the correlation between the three scale variables and the one-year status

Table 1

Scale variable	Eta test	Eta-squared	Sig
AGE	0.039	0.002	0.400
Volume	0. 043	0.002	0.345
Forced capacity	0.046	.002	0.316

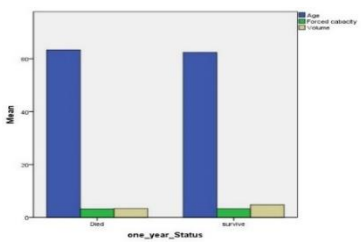


Fig 1

Step 3-Correlation between the nominal variables and one-year status: The Chi-squared test (Mchugh, 2013) was used to determine the correlation between 13 nominal variables and the one-year status.

Table 2

Nominal variables	Values	Asymp. Sig. (2-sided)
Diagnosis	19.336	0.002
Performance_status	4.480	0.5
Pain_before_surgery	1.547	0.6
Haemoptysis_before_surgery	2.034	0.34
Dyspnoea_before_surgery	5.234	0.23
Cough_before_surgery	3.711	0.33
Weakness_before_surgery	3.541	0.43
Size_of the original tumour	16.570	0.003
Type _2 DM - diabetes mellitus	5.581	0.002
MI up to 6 months	0.351	0.5
PAD - peripheral arterial diseases	0.656	0.45
Smoking	3.473	0.32
Asthma	0.351	0.45

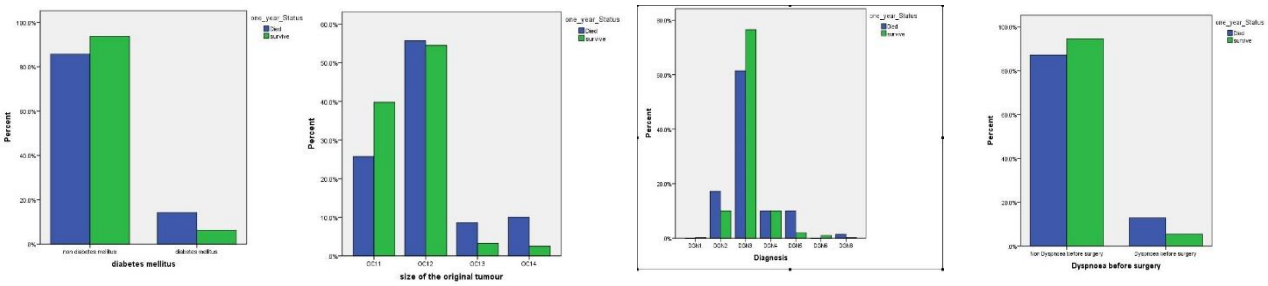


Figure 2

FINDINGS

Table.1 shows the values of the three scale variables associated with the nominal variable the one-year status. We used (α=0.05). From the tables, we can see that the three values have large significant values than 0.05. thus, no association was found between this correlation. Figure 1 shows, the independence of the three variables from the one-year status. Table 2 represents the values of the chi-square test as well as the significant value. We used (α=0.05). four variables only show significant association with one status year. Figure 2 shows the percentage of the nominal variables and one-year status.

IMPLICATIONS

In this study, we used statistical tests to analyze and visualize the **Thoracic Surgery Data**. We used two types of relations. The Eta test between the one-year status and nominal variables while the chi-square between one-year status and the nominal variables. From the 16 variables, the related variables whit the one year-status were only four variables. Therefore, we could create a model with less number of variables by understanding and visualizing the dataset.

REFERENCES

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 9: One-way analysis of variance. *Critical Care*, *8*(2), 130–136.

Bilbao, I., & Bilbao, J. (2018). Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks. *2017 IEEE 8th International Conference on Intelligent Computing and Information Systems, ICICIS 2017*, *2018–January*(Icicis), 173–177.

Mchugh, M. L. (2013). The Chi-square test of independence Lessons in biostatistics. *Biochemia Medica*, *23*(2), 143–9.

Mirkin, B. (2011). Data Analysis , Mathematical Statistics , Machine. *ICACSIS*, 978–979.

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing Journal*, *14*(PART A), 99–108.

Trends in Higher Education: Intelligence Amplification and Learning Analytics

Gan Chin Lay and Liew Tze Wei
Faculty of Business, Multimedia University, 75450 Melaka, Malaysia

Corresponding Email: gan.chin.lay@mmu.edu.my

Keywords: *Student-Centered Learning, Intelligence Amplification, Learning Analytics, Artificial Intelligence, Higher Education*

Global higher education, including in Malaysia, is facing rapid changes. Scholars in recent years are actively advocating for tertiary institutions to foster student-centered learning environment where students are encouraged to be independent self-directed learners (Robinson et al., 2016). Though predominantly associated with e-learning, a growing awareness and recognition of the benefits of student-centered learning has brought it into face-to-face classrooms as well. That being said, fostering independent self-directed learning amongst Malaysian tertiary students is by no means an easy task after years of teacher-led schooling. However, technological advancements are set to play a major role in supporting student-centered learning experience for both students and educators of tertiary institutions.

Firstly, the effective use of information technology to augment human intelligence, known as Intelligence Amplification (IA), an idea that was proposed in the 1950s has resurfaced largely due to strides made in Artificial Intelligence (AI) in areas such as machine learning, natural language processing, and computer vision. The argument on whether AI actually promotes IA is debatable. Whether one agrees or disagrees, AI's influence and reach in education is pervasive, from dynamic assessments and feedback on students' submissions; intelligent tutoring systems for personalized learning support; educational chatbots, i.e. automated intelligent agents that are able to provide answers to students' questions; to enabling immersive learning using augmented reality (Schmidt, 2017). Utilizing these technologies effectively are key for supporting IA, the ultimate goal in all teaching and learning endeavours.

Despite the sophistication of these emerging technologies, understanding learning behaviours is imperative. The process of analysing large amount of data in learning behavioural analyses, and the reporting of data about the development and performance of learners are known as learning analytics (Sclater et al., 2016). Sophisticated data-mining techniques and statistical modelling, big data, efficient and powerful analytical software are poised to provide answers to these key questions: how does learning happen, what drives learning, what are the essentials for improving learning experiences and practices, what are the strategies for improving engagement and retention rates, and lastly, how can we reduce and ultimately prevent ineffective learning from taking place using predictive analytics. It is opined that learning analytics and IA are interrelated. Knowing the answers to these questions are deem crucial for developing educational technologies that are able to support and foster student-centered learning in higher education.

REFERENCES

Robinson, S., Neergaard, H., Tanggaard, L., & Krueger, N. F. (2016). New horizons in entrepreneurship education: from teacher-led to student-centered learning. *Education+ Training*, *58*(7/8), 661-683.

Schmidt, A. (2017, December 37). How AI Impacts Education. Retrieved from https://www.forbes.com/sites/theyec/2017/12/27/how-ai-impacts-education/#4f118a1b792e

Sclater, N., Peasgood, A., & Mullan, J. (2016, April 19). Learning analytics in higher education: A review of UK and international practice. Retrieved from https://www.jisc.ac.uk/reports/learning-analytics-in-higher-education

The Development of a Conceptual University Student Cybersecurity Behavioral Model (C-Uscb) Based on the Impact of Multiple Factors and Constructs of Self-Reported Cybersecurity Behaviors

Fatokun Faith Boluwatife, Suraya Hamid and Azah Norman
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Emails: evangfatoks@gmail.com, suraya_hamid@um.edu.my

Keywords: *Cybersecurity, Cybersecurity behaviors, University Students Cyber-behaviors, Human behaviors*

INTRODUCTION

Due to the fact that the access to internet is swiftly escalating across the globe, vis-à-vis the expansion of larger connectedness across individuals, finance, and business, building necessary safeguards against privacy and security will only be of more importance. This actuality therefore makes cybersecurity, as well as other outstanding online and computer safeguarding practices to be the major critical problems of our generation. Coming to the aspect of cybersecurity behaviors, as informed by (M. Gratian et al., 2018), and other researchers, it is clear that humans are identified as the major weak link in cybersecurity. This is due to the fact that many technical security solutions still prove worthless and not effective enough because of human-made mistakes and negligence's of behavior while surfing the internet. It is in view of this concern that researchers such as (Anwar et al., 2017; Halevi et al., 2016) and many others have carried out research in different forms and approaches with regards to cybersecurity behaviors among humans. More importantly, if humans' online behavior improves, then they will have better cybersecurity assurance.

PROBLEM STATEMENT

Cybersecurity threats are rampant everywhere on the cyberspace, even though several technologies are in place to combat such attacks, the offenders keep strategizing and unfortunately rely mostly on the careless behaviors of users online to attack them. Furthermore, many of the research that have been conducted to measure cybersecurity human behavioral impacts have focused more on organizational staff, thereby leaving the ordinary cyber users out of the picture. It is in view of this gap that this research tends to extend an investigative quantitative research approach through the use of some adapted cybersecurity behavioral constructs to measure the impact of self-reported cybersecurity behaviors among university students in particular. This investigation will then end up in the development of a conceptual university student cybersecurity behavioral model(C-USCB). The purpose is to discover if there are significant differences between the different factors that will be used coupled with the adapted constructs of self-reported cybersecurity behaviors.

METHODOLOGY

This research shall employ a quantitative research approach. The participants shall be students of a particular university in Malaysia. A survey-based investigation shall be conducted on them with regards to their behaviors/attitudes towards cybersecurity issues. We shall send out questionnaires via an online medium to collect primary data, also an alternative preparation will be made for paper-based questionnaires in case we experience any delay in data collection process. The questions that shall be used as measurement tools in this investigation have been adapted from constructs derived from the health belief Model by (Becker et al., 1978) and the Protection Motivation Theory by (Maddux & Rogers, 1983). More specifically, these constructs are: Self-Reported Cybersecurity Behavior, Security Self-Efficacy, Peer-Behavior, Information-Seeking Skills Using the Internet, Internet Skills, Computer Skills, Prior Experience with Computer Security Practices and Perceived Vulnerability. Also demographic details of the students shall also be collected and tested on, such

as age, gender, major, level of education. Analysis such as T-test, ANOVA, Correlation, and Regression test shall be performed using SPSS package while the Research Conceptual Model shall be developed using the SMART-PLS (SEM) software.

FINDINGS

From the literature review conducted so far, a brief summarized findings based on different authors research focus, and method used, have been coupled together in a tabular form below:

Table 1: Lit Review brief findings based on Research Focus and Approach

Research Focus	Method/Approach used	Reference
Organization Workers	Quantitative	(Anwar et al., 2017; Hadlington, 2017; Hadlington & Murphy, 2018; N. S. Safa et al., 2016; Vance et al., 2012)
	Qualitative/Observational	
	Mixed	(Nader Sohrabi Safa et al., 2015)
Personal Users	Quantitative	(Halevi et al., 2016; Ogutcu et al., 2016)
	Qualitative/Observational	
	Mixed	(Egelman et al., 2016; Rajivan et al., 2017)
University Students	Quantitative	(Margaret Gratian et al., 2018; Yan et al., 2018)
	Qualitative/Observational	
	Mixed	

From the table, bulk of research focused on Organization workers and just a little on both personal users and university students, this shows the need for this current research. Also, the quantitative method is the most used methods, showing that it is the most suitable way of conducting such research since it deals with proven measurement constructs. Some Cybersecurity behavioral studies have measured on different factors and have developed different models, some of them are stated as follows: (Anwar et al., 2017) focused on gender difference and employees cybersecurity behaviors and discovered that gender had effects on the cybersecurity behavior of employees, (Margaret Gratian et al., 2018) study on human traits and cybersecurity, found out that 5-23% of the variance in cybersecurity behaviors was based on personal differences, and (Halevi et al., 2016) focused on cultural and psychological factors in cybersecurity and found out that cybersecurity behaviors differs across different cultures and countries.

IMPLICATIONS

This research will be of much significance by adding to the body of knowledge, it will also be useful to Industry practitioners in order to enlighten them on redesigning their security software with the human factor in mind, and it will be of utmost importance to the university community as well especially for students or researchers interested in this domain. Finally, it will educate the participants by cautioning them of their security behaviors on the internet.

REFERENCES

Anwar, M., He, W., Ash, I., Yuan, X., Li, L., & Xu, L. (2017). Gender difference and employees' cybersecurity behaviors. *Computers in Human Behavior, 69*, 437-443. doi: 10.1016/j.chb.2016.12.040

Becker, M. H., Radius, S. M., Rosenstock, I. M., Drachman, R. H., Schuberth, K. C., & Teets, K. C. (1978). Compliance with a medical regimen for asthma: a test of the health belief model. *Public Health Reports, 93*(3), 268-277.

Egelman, S., Harbach, M., & Peer, E. (2016). *Behavior ever follows intention?: A validation of the security behavior intentions scale (SeBIS).* Paper presented at the Proceedings of the 2016 CHI conference on human factors in computing systems.

Gratian, M., Bandi, S., Cukier, M., Dykstra, J., & Ginther, A. (2018). Correlating human traits and cyber security behavior intentions. *Computers and Security, 73*, 345-358. doi: 10.1016/j.cose.2017.11.015

Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. *Heliyon, 3*(7), e00346.

Hadlington, L., & Murphy, K. (2018). Is Media Multitasking Good for Cybersecurity? Exploring the Relationship Between Media Multitasking and Everyday Cognitive Failures on Self-Reported Risky Cybersecurity Behaviors. *Cyberpsychology, Behavior, and Social Networking, 21*(3), 168-172.

Halevi, T., Memon, N., Lewis, J., Kumaraguru, P., Arora, S., Dagar, N., . . . Chen, J. (2016). *Cultural and psychological factors in cyber-security*. Paper presented at the Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, Singapore, Singapore.

Maddux, J. E., & Rogers, R. W. (1983). Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology, 19*(5), 469-479. doi: 10.1016/0022-1031(83)90023-9

Ogutcu, G., Tastik, O. M., & Chouseinoglou, O. (2016). Analysis of personal information security behavior and awareness. *Computers & Security, 56*, 83-93. doi: 10.1016/j.cose.2015.10.002

Rajivan, P., Moriano, P., Kelley, T., & Camp, L. J. (2017). Factors in an end user security expertise instrument. *Information and Computer Security, 25*(2), 190-205. doi: 10.1108/ics-04-2017-0020

Safa, N. S., Solms, R. V., & Futcher, L. (2016). Human aspects of information security in organisations. *Computer Fraud and Security, 2016*(2), 15-18. doi: 10.1016/S1361-3723(16)30017-3

Safa, N. S., Sookhak, M., Von Solms, R., Furnell, S., Ghani, N. A., & Herawan, T. (2015). Information security conscious care behaviour formation in organizations. *Computers & Security, 53*, 65-78.

Vance, A., Siponen, M., & Pahnila, S. (2012). Motivating IS security compliance: insights from habit and protection motivation theory. *Information & Management, 49*(3-4), 190-198.

Yan, Z., Robertson, T., Yan, R., Park, S. Y., Bordoff, S., Chen, Q., & Sprissler, E. (2018). Finding the weakest links in the weakest link: How well do undergraduate students make cybersecurity judgment? *Computers in Human Behavior, 84*, 375-382.

Analyzing and Visualizing Data Dengue Hotspot Location

Nadzurah Binti Zainal Abidin and Amelia Ritahani Ismail
Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

Corresponding Emails: nadzurah.zabidin@gmail.com, amelia@iium.edu.my

Keywords: *Correlation, Visualization, Machine Learning, Naïve Bayes.*

INTRODUCTION

In this paper, we will explore the Dengue Hotspot Location training data set that publicly available at data.gov.my. The data set consists of 10,116 cases reported according to respective district in Malaysia for 5 years, starting from 2011 until 2015. The dataset contain 7 columns which are: Tahun, Minggu, Negeri, Daerah/Zon, Lokaliti, Jumlah Kes Terkumpul, and Tempoh Wabak Berlaku (Hari). The purpose of this study is to measure strength of the correlation between all variables in dataset Dengue Hotspot Location. This paper also focused primarily on the selection of suitable variables from a large data set and imputation of missing values. Many statistical models has proven to be fail with missing values. Besides, many researchers had proposed various ways to handle missing values. However, in this paper we demonstrate our approach for analyzing data with one of the machine learning classifier, Naïve Bayes. The choices were made from the highest accuracy among four machine learning classifiers experimented in the previous paper (Abidin, Ritahani, & Emran, 2018).

PROBLEM STATEMENT

Before moving to the critical part in analyzing data, we should study the data and find correlation between each variables. This is due to the fact that oddities in the data can cause bugs and muddle the results. One of the common issues to cater in this paper is to handle missing values. Missing values may results in less accurate estimations and reduced the quality of data set.

METHODOLOGY

The first step of data analysis or predictive modeling activities is an initial exploration of data. The step in exploratory analysis is reading, understanding the data, and then exploring the variables. The initial exploration of data began with exploring the numerical variables for min, max, mean and median as Table 1 below.

Table 1: Numerical Variables Summary

Column Name	Min	Max	Mean	Median
Tahun	2011	2015	2014	2014
Minggu	1.00	53.00	23.79	20.00
Jumlah Kes Terkumpul	2.00	833.00	23.67	15.00
Tempoh Wabak Berlaku Hari	1.00	387.00	59.75	49.00

While Table 2 shows attribute information for categorical variables with description.

Table 2: Categorical Attributes Information

Column Name	Description
Negeri	States in Malaysia; eg Selangor, Johor, Kelantan, Negeri Sembilan, Pulau Pinang, Perak, Sabah, Sarawak, and Terengganu.
Daerah	District in Malaysia. Approximately 70 district mentioned.
Lokaliti	Name of places in respective district.

Secondly, before the real experiment began, we visualize data set in a meaningful manner using R. The statistical model will be presented as we shall see in the next section. After modeling data set, the essential part conducted is feature selection. Feature selection is a process of removing redundant, inconsistencies and impute missing values. We proposed to use naïve bayes as our approach to impute missing values.

Next step covers on measuring the correlations of each variables and identify the correlation relationship and prepare new model using spearman correlation.

FINDINGS

According to the result below, there is a significant correlation between minggu, jumlah kes terkumpul, and tempoh wabak berlaku. The figure 1 shows that there a strong correlation between minggu and tempoh wabak berlaku, while there are significant correlation between tahun and jumlah kes terkumpul. Nevertheless, the highest correlation can be seen between jumlah kes terkumpul and tempoh wabak berlaku. The correlation using Spearman shows a significant results after numerical variables has been imputed.

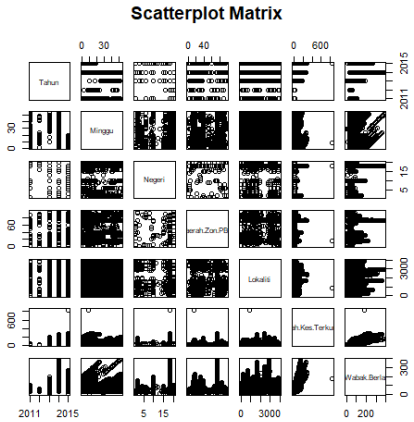


Figure 2: Scatterplot Matrix Dengue Hotspot Location

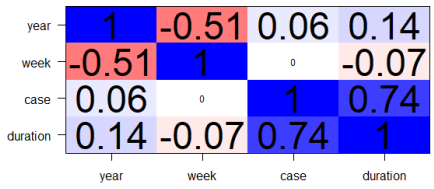


Figure 2 (a): Spearman Correlation before Impute

Figure 3 (b): Spearman Correlation after Impute

The result drawn from the figure 2 (a) and (b) concludes that there are slightly changes correlation after imputation between all numerical variables. The changes were not much since the percentage of missing value of Dengue Hotspot Location is small with only 1.97%.

IMPLICATIONS

The result shows that the dengue cases in Malaysia are expected to increase steadily between ends of the year to early of the following year. The major contribution to this prediction is dengue cases likely to rise during monsoon season, where Malaysia's monsoon season falls from October to March.

REFERENCES

Abidin, N. Z., Ritahani, A., & Emran, N. A. (2018). Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(6).

https://doi.org/10.14569/IJACSA.2018.090660

Causal Discovery of Gene Regulatory Networks (Grns) from Gene Perturbation Experiment

Winddy Pindah[1], Sharifallilah Nordin[2] and Ali Seman[3]
Universiti Teknologi MARA (UiTM)

Corresponding Email: shendy1015@gmail.com, sharifa@tmsk.uitm.edu.my, aliseman@tmsk.edu.my

Keywords: *Reconstruction of gene regulatory networks (GRNs), Causal Inference, Causal Discovery, Bayesian Network, Identifiability, MCMC algorithm*

INTRODUCTION

In most applications, the gene regulatory networks (GRNs) structure is unknown and has to use reverse engineering procedure (Tegner, Yeung, Hasty, & Collins, 2003) to reveal network topologies and regulatory interactions among the genes in a living cell from gene expression data. The biochemical process involves in gene regulation can be represented as directed edges in GRNs which correspond to the causal (or cause-effect) relationship among genes. The causal relationship between two genes, gene $X \rightarrow$ gene Y, can be represented as a directed acyclic graph (DAG). The recent emergence of multiple technologies for gene perturbation experiment provides us data to infer causal relationship among genes in GRNs. Reverse engineering of GRNs or also refers to the reconstruction of GRNs from gene perturbation experiment rely on comparison between the expression values of various genes in wild type cells (observational data) and knock-out or knock-down cells (interventional data); the underlying idea is that if gene Y is regulated by gene X, then its expression values under a knock-out of gene X will be different from the value in a wild type experiment and gene X must a cause of change in gene Y. This strategy essentially follows the logic of standard causal discovery from mixed observational and interventional data (Das, Caragea, Welch, & Hsu, 2010).

PROBLEM STATEMENT

Causal discovery aims to discover causal structure in form of DAG from a set of random variables to derive causal model. In this case, the GRNs we refer as a causal model. A causal model is a Bayesian network whose edge whose edges has causal significance (Pearl, 1988). Given a causal model allows us to perform causal inference, for example, estimating the causal effect of intervention of some genes. However, many of the causal discovery algorithm are plagued by fundamental identifiability issues, that is, incorrectly infer the causal relationship between two variables from given data (Hauser & Buhlmann, 2015; Meinshausen et al., 2016). Moreover, even with a sufficient interventional data from gene perturbation experiment, i.e. roughly one knock-out for each gene, it is still often not possible to identify a unique causal structure from given data (Marbach et al., 2010). Perhaps due to may be due to the heterogeneous coverage of the gene network space (Altay et al., 2010). As such, a number of works use marginal distribution over the intervention variable in the model to estimate their causal effects rather than attempting to use the fully specified causal model (Daphne Koller & Friedman, 2009; Maathuis, Kalisch, & Bühlmann, 2009; Monneret, Jaffrézic, Rau, Zerjal, & Nuel, 2017; Tian & Pearl, 2001). This approach, however, not always biologically relevant.

METHODOLOGY

A Markov chain Monte Carlo (MCMC) algorithm based on random walks is a typical causal discovery algorithm used to explore the search space and select the highest scoring causal structure from mixed observational and intervention data (Friedman & Koller, 2003; Madigan & York, 1993). In order to address the problem of identifiability of unique causal structure from mixed observational and interventional data, we propose a new framework for causal discovery based on the strong monotonic effects and weak monotonic effects exists in the causal relationship among nodes in casual structure

(Vanderweele & Robins, 2010; VanderWeele & Robins, 2009). This framework builds on the use of MCMC algorithm. In this idea, there is a presence of true causal relationship between gene X and gene Y if gene X has a strong monotonic effect on gene Y than gene X has a weak positive monotonic effect on gene Y. Otherwise, the two genes have no true causal relationship.

FINDINGS

The propose framework for causal discovery based on use of MCMC algorithm expected to infer the presence of a true causal effect even in the presence of unmeasured confounding from mixed observational and interventional data.

IMPLICATION

The finding of this study can be used to identify more accurate gene regulatory networks (GRNs) from gene perturbation experiment data. In the same time can contribute to the improvement performance causal discovery algorithm. Causal discovery is the goal for scientific exploration, and causal discovery in data is what computational researchers are able to contribute greatly to our society.

REFERENCES

Daphne Koller, & Friedman, N. (2009). *Probabilistic Graphical Models: Principle and Techniques*.

Das, S., Caragea, D., Welch, S. M., & Hsu, W. H. (2010). *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*. *Medical Information Science Reference*. Hershey, New York. https://doi.org/10.4018/978-1-60566-685-3

Friedman, N., & Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery. *Machine Learning*, *50*, 95–125.

Hauser, A., & Buhlmann, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *77*(1), 291–318. https://doi.org/10.1111/rssb.12071

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, *37*(6 A), 3133–3164. https://doi.org/10.1214/09-AOS685

Madigan, D., & York, J. (1993). *Bayesian Graphical Models for Discrete Data*.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., & Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *PNAS*, *113*(27), 7361–7368. https://doi.org/10.1073/pnas.1510493113

Monneret, G., Jaffrézic, F., Rau, A., Zerjal, T., & Nuel, G. (2017). Identification of marginal causal relationships in gene networks from observational and interventional expression data. *PLoS ONE*, *12*(3), 1–13. https://doi.org/10.1371/journal.pone.0171142

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference* (First Edit). Morgan Kaufmann.

Tegner, J., Yeung, M. K. S., Hasty, J., & Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, *100*(10), 5944–5949. https://doi.org/10.1073/pnas.0933416100

Tian, J., & Pearl, J. (2001). Causal Discovery from Changes. *Uai*, 512–521. https://doi.org/10.1.1.20.7579

Vanderweele, T. J., & Robins, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *72*(1), 111–127. https://doi.org/10.1111/j.1467-9868.2009.00728.x

VanderWeele, T. J., & Robins, J. M. (2009). Properties of Monotonic Effects on Directed Acyclic Graphs. *Journal of Machine Learning Research*, *10*, 699–718.

A Deep Convolutional Neural Networks on Malaysian Food Classification

J. Joshua Thomas and Naris Pillai
Department of Computing, School of Engineering, Computing and Built Environment, KDU Penang University College, Penang Malaysia

Corresponding Emails: joshopever@yahoo.com, Narispillai@gmail.com

INTRODUCTION

Predictable machine-learning systems were constrained in their capacity to process normal data in their raw shape. For a considerable length of time, developing pattern-recognition or machine-learning framework required cautious building and extensive area aptitude to outline a component extractor that changed the crude information, (for example, the pixel estimations of a picture) into an appropriate interior portrayal or feature vector from which the learning subsystem, frequently a classifier, could identify patterns from the input (De Sousa Ribeiro, F et al., 2018).

Representation learning is an arrangement of strategies that enables a machine to be nourished with crude data and to consequently find the representations required for detect or classify. Deep-learning strategies (Szegedy, C et al., 2015) are portrayal learning techniques with different levels of representation, acquired by creating straightforward yet non-direct modules that each change the portrayal at one level (beginning with the raw data) into a representation at a higher, somewhat more dynamic level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by recognizing particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by engineers: they are learned from data using a general-purpose learning procedure.

METHODOLOGY

Image classification is the task of assigning a single label to an image (or rather an array of pixels that represents an image) from a fixed set of categories. A complete pipeline for this task is as follows:

- **Input:** A set of N images, each labeled with one of K different classes. This data is referred to as the training set.
- **Learning** (aka Training): Use the training set to learn the characteristics of each class. The output of this step is a model which will be used for making predictions.
- **Evaluation**: Evaluate the quality of the model by asking it to make predictions on a new set of images that it has not seen before (also referred to as the test set). This evaluation is done by comparing the true labels (aka ground truth) of the test set with the predicted labels output by the learned model.

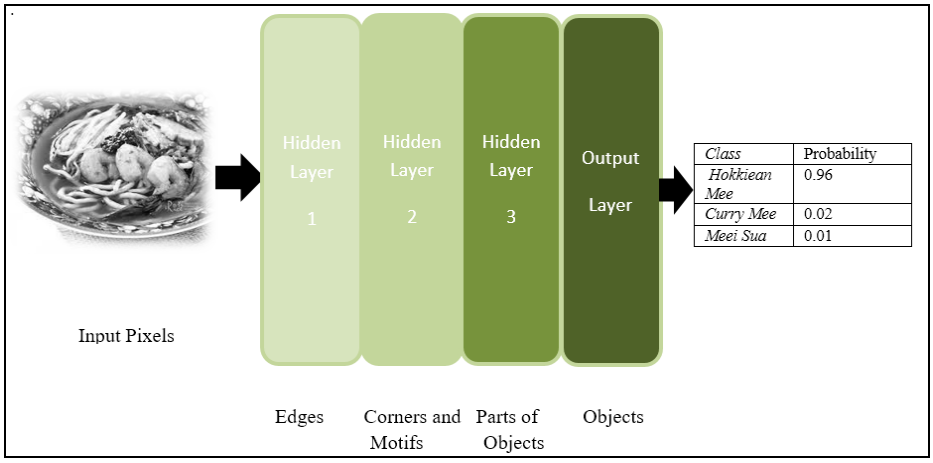


Figure 1: Insertion model of multilayer perceptron

The ultimate goal of our model is to generate descriptions of image regions. During training, the input to our model is a set of images and their corresponding sentence descriptions (Figure 1). The propose model that aligns sentence snippets to the visual regions that they describe through an embedding. Then treat these correspondences as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets.

DATA COLLECTION AND PRE-PROCESSING

The collected our dataset (5115 images as dated 30/4/2018) using the Google Image Search, Bing Image Search (search engines) API and on the spot image collection at the restaurants, Hawkers (Street food) stalls and street food. The work has explored the use of ImageNet (imagenet_cvpr,2009) and (Flickr) for collecting images. However, the images from Google and Bing to be much more representative of the classes they belonged to, compared to the images from ImageNet and Flickr. ImageNet and Flickr seem to have a lot of spurious images (images which clearly do not belong to the class). Hence decided to use the images could collect from Google and Bing.

EXPERIMENTAL RESULTS

The results of the proposed work will analyses the CNN architectures on how accurate in solving the complete the classification while training the multiclass datasets. The working nature of the algorithm as compared with the classification accuracy and the number of iterations (Epoch) for the fully connected two layers are evaluated. While the increase of dataset will be trained, there will be a parameter adjustment by using restricted Boltzmann machine (RBM) are to be analyzed for the accuracy of the proposed training and test model with DNN architectures.

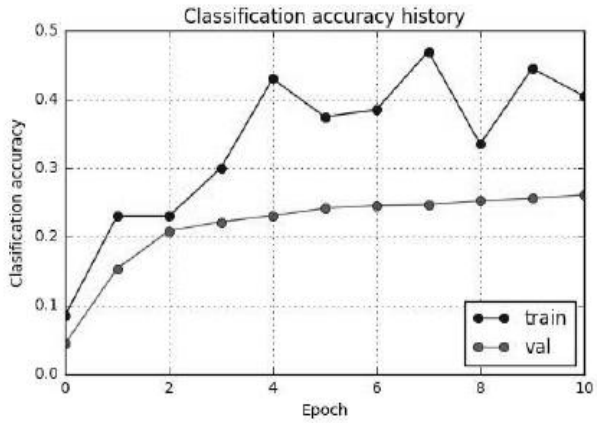


Figure 2: classification accuracy history of a fully connected two layer neural network using image features.

After the five conv layers, we added two fully connected layers with 5119 and 14 neurons respectively. For the last layer we use *softmax* with cross entropy Loss. The best validation accuracy of 0.40 was achieved using the Adam (Li, L. J., Socher, R., & Fei-Fei, L. (2009, June) update rule with a learning rate of 1e-04. The test set accuracy was 0.40. The results could be improved once the train model has include with additional and hidden layers and the input image size with good resolutions.

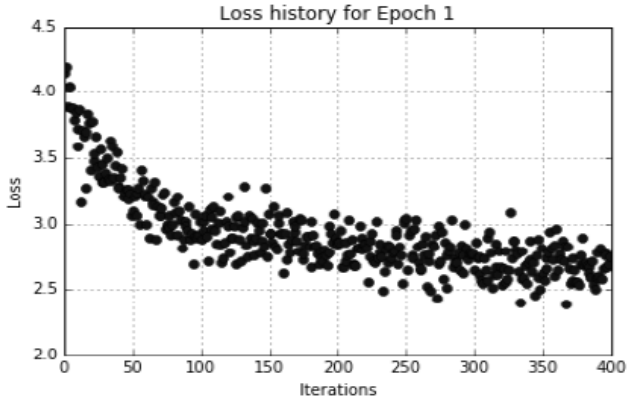


Figure 3 Reduction in loss in accuracy first epoch of conventional network.

DISCUSSION

As shown in Figure 2, and Figure 3 the prediction accuracy increased with the increase number of image inputs. The results could be improved with a novel type of learning machine, called support vector machine (SVM), has been receiving increasing attention in areas ranging from its original application in pattern recognition to the extended applications. SV M has the greater generalization ability over neural works. The work will be extended with SVM over Deep Neural Network to improve the training accuracy and entropy loss. The results in this work has attributable to the fact that the implementation of convolutional neural network ability to classify the images with pattern recognition with appropriate accuracy of the training models.

REFERENCES

De Sousa Ribeiro, F., Caliva, F., Swainson, M., Gudmundsson, K., Leontidis, G., & Kollias, S. (2018, May). An adaptable deep learning system for optical character verification in retail food packaging. Evolving and Adaptive Intelligent Systems, IEEE Conference on..
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Li, L. J., Socher, R., & Fei-Fei, L. (2009, June). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2036-2043). IEEE.

Arabic Sentiment Analysis: An Overview of the ML Algorithms

Mohamed Elhag M. Abo[1], Nordiana Ahmed[2] and Vimala Balakrishnan[1]
[1]Department of Information Systems Faculty of Computer Science and Information Technology, University of Malaya
[2]Department of Library Science & Information Faculty of Computer Science and Information Technology, University of Malaya

Corresponding Emails: aboo72me@gmail.com, vimala.balakrishnan@um.edu.my

INTRODUCTION

The Machine Learning (ML) algorithms is a powerful technique for sentiment analysis, merely having a machine learning algorithm is not sufficient for the Arabic language sentiment analysis without proper pre-processing steps (Duwairi & El-Orfali, 2014). Moreover, the algorithm needs to evaluate by applying it to standard Arabic standard data set to choice best model and ML algorithms preferences and ultimately provide them with a customised experience. However, there is no standard and appropriate dataset available for Arabic the overall evaluation of ML algorithms (Zaidan & Callison-Burch, 2014).

Notwithstanding, many machine learning algorithms give promising results with the use of appropriate pre-processing. In this paper, we perform an extensive survey to identify the most used of ML algorithms for Modern Standard Arabic (MSA) and Dialect Arabic (DA) from the literature.

METHODOLOGY

In an attempt to perform an exhaustive search, the most accurate and reliable bibliographic databases that cover the most important journal articles and conference proceedings are identified. Moreover, the total number of sentiment analysis papers published in four databases Association for Computing Machinery (ACM), ScienceDirect (SD), IEEE Xplore (IEEE), and the web of science (WoS). Furthermore, the article published is 1458, with Sentiment analysis and only 48 published on Arabic sentiment analysis. Moreover, the study was done between 2011 to 2017. However, Arabic sentiment analysis needs more research.

FINDINGS

Based on the reviewed more than 87 essential journal articles and conference, the most ML algorithms used for MSA and DA is SVM as shown in Fig 1 below.

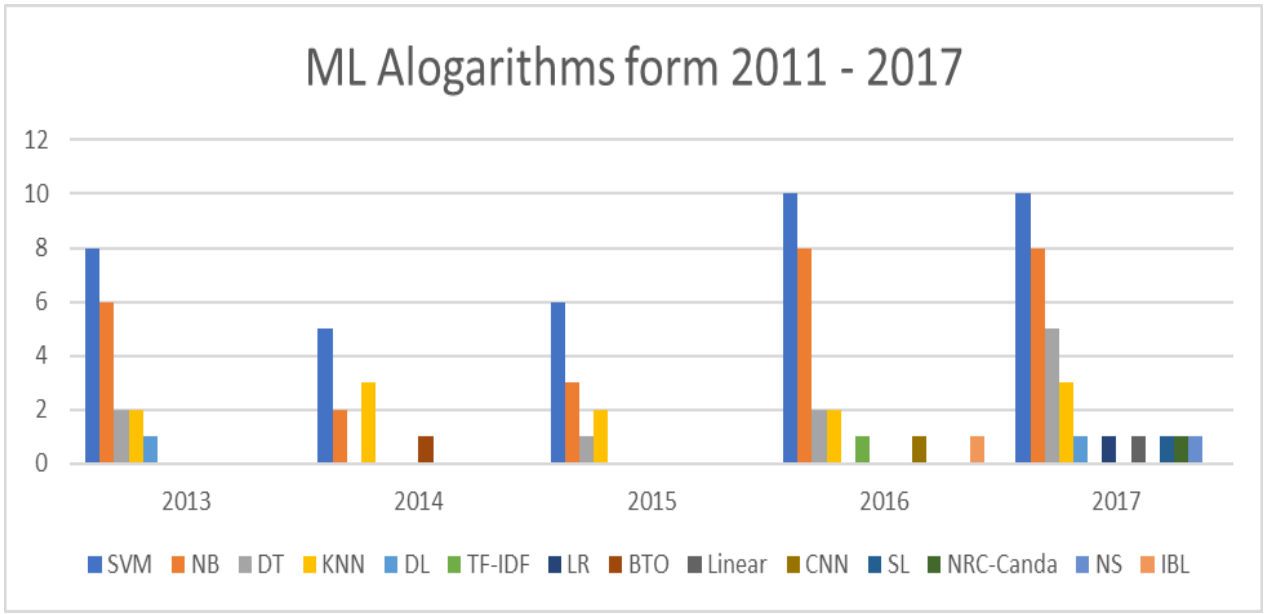


Table 1: Summary of Most ML algorithms and result for Arabic sentiment analysis

Year	Language	Approach	Algorithms	Result
2017	Saudi dialect Arabic	lexicon-based	weighted lexicon-based algorithm (WLBA)	WLBA Without Do'aa rules accuracy = 77.6% With Do'aa rules accuracy = 85.4% (P and N Do'aa) Performance of WLBA Saudi Twitter Dataset accuracy = 81% El-Beltagy and Ali [18] accuracy =76 %
2017	Modern Standard Arabic and Iraqi, Egyptian and Lebanese dialects Arabic	supervised	K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision tree (DT)	MSA NB Precision = 78.81% recall =93.94% f-measure =82.58 % KNN Precision = 82.12% recall = 86.83% f-measure =82.90% DT Precision =82.74% recall =92.14 % f-measure =85.54% DA NB Precision =86.76 % recall = 92.52% f-measure = 87.97% KNN Precision = 87.45% recall = 88.77% f-measure = 87.22% DT Precision = 89.53% recall = 87.52% f-measure =86.78%
2017	Modern Standard Arabic	supervised	Naive Bayes (NB), LR, Support Vector Machines (SVM), DNNs and CNNs "Unigram", "Bigram", TF-IDF	Multinomial Naive Bayes accuracy = 90.14% Bernoulli Naive Bayes accuracy = 90.14% Logistic Regression accuracy = 86.94% Support Vector accuracy = 90.88% Linear Support Vector accuracy = 91.37% Stochastic Gradient Descent accuracy =91.87% Nu-Support Vector accuracy = 87.82% Deep Neural Network reached about 85%
2016	Modern Standard Arabic	Supervised	Naive Bayes (NB)	NB classifier accuracy = 90%
2016	Modern Standard Arabic	supervised	Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN)	SVM Precision = 0.948% and recall =0.939% NB Precision = 0.946% and recall = 0.939% KNN Precision = 0.803% and recall = 0.776%

The findings of the work presented in table 1 show SVM and NB is most used in the study of Arabic sentiment analysis. Moreover, other algorithms are also got a good result such as accuracy of Stochastic Gradient Descent is got =91.87%. The evaluation of the ML algorithms performance done by a set of metrics such as accuracy, f-measure, recall and precision. However, some algorithms take a long time and memory consumption.

CONCLUSION

Machine learning algorithms for modern standard Arabic need an excellent preprocessing to give a good result. The studies were shown no doubt that there is a significant problem in choosing the right machine learning algorithms for the Arabic language. Furthermore, depends on building the structure of the analysis model from the first steps in the right way, such as the preprocessing, selection of the ML algorithms and then choosing the correct valuation metrics. This paper performs an extensive survey to identify the most used machine learning algorithms in MSA. Moreover, one important observation is that there are unavailable identified standard data sets are available for the Arabic sentiment analysis.

REFERENCES

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, *40*(4), 501-513.

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic Dialect Identification. *Computational linguistics*, *40*(1), 171-202. doi:10.1162/COLI_a_00169

Feature Selection for Heart Disease Prediction

Nashreen Md Idrs[1], Chiam Yin Kia[1], Kasturi Dewi Varathan[2], Lau Wei Tiong[2]
[1]Deparment of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya
[2]Deparment of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya

Corresponding Emails: nashreen.idris@gmail.com, yinkia@um.edu.my, kasturi@um.edu.my, weitiong@gmail.com

Keywords: *Heart disease; cardiovascular disease; Feature selection method; Predictio*

INTRODUCTION

Rajeswari et al (2012) said that cardiovascular disease (CVD) will be the largest cause of death and disability by 2020 in India. In 2020, 2.6 million Indians are predicted to die due to coronary heart disease which constitutes 54.1 % of all CVD deaths. The possible of data mining to improve most of problems involve in medical sector already being detected as early by 1997 by the World Health Organization (WHO) (Gulbinat, 1997). The WHO also mentioned that medical diagnosis and prediction will give the opportunities as the application of knowledge detection in medical data repositories as the whole. Techniques used in data mining can help most of decision making developing for the prediction in diseases that using multiple sets of medical datasets (Nilashi, bin Ibrahim, Ahmadi, & Shahmoradi, 2017). The deaths occur due to lack of early medical diagnosis of cardiovascular disease, thereby posing a big challenge to health care organizations. Accurate and timely diagnosis of patients is required for effective treatment and for quality service (Narain, Saxena, & Goyal, 2016). By implementing the feature selection method, it will improvise the generalize capabilities and reduces complexity and execution time. These will help to give the prediction of medical dataset as quickly as possible (Shilaskar & Ghatol, 2013).

MATERIALS AND METHOD

The experiment is executed using Jupyter Notebook in Python language and the dataset from UCI machine learning repository specifically Cleveland Heart Disease is used to evaluate the prediction accuracy. The data is cleaned first to remove inaccurate or irregular data in order to ensure its quality. There are three feature selection method chosen to run this experiment are Univariate Selection (US), Recursive Feature Elimination(RFE) and Feature Importance(FI).

RESULTS AND DISCUSSION

In this section, among 14 attributes available in the dataset, only the top five features selected from every method is tabulated below.

Table 1 Top five features from every method

No	Method	Top Five Features
1	Univariate Analysis	thalach, ca, oldpeak, thal, exang
2	Recursive Feature Elimination	fbs, oldpeak, ca, sex, exang
3	Feature Importance	thalach, ca, thal, oldpeak, age

Table 2 Abbreviation description

No	Abbreviation	Description
1	thalach	Maximum heart rate achieved
2	ca	Number of major vessels (0-3) colored by flourosopy
3	oldpeak	ST depression induced by exercise relative to rest
4	thal	Thal
5	exang	Exercise induced angina
6	fbs	Fasting blood sugar > 120 mg/dl
7	sex	Sex

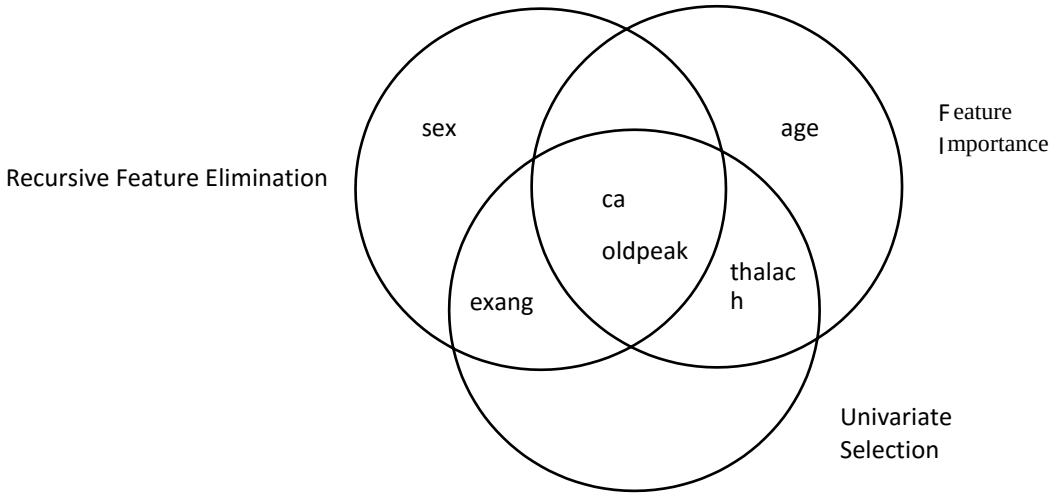


Figure 1 Subset of top five features among three feature selection methods

Based on the results, we can observe that every method selects different features as their top features. However, all the three method has two common top feature which is oldpeak and ca. The common top features between US and FI are thalach and thal while the common top feature between US and RFE is exang. But there are no common top feature between FI and RFE. To conclude, the top five significant features for heart disease prediction are ca, oldpeak, thalach, thal and exang.

CONCLUSION

Data mining techniques can be used to analyse the raw data to provide new insights towards the goal of disease prevention with accurate predictions and by implementing feature selection method, the prediction accuracy increased higher. Heart disease is one of the main causes of death in this world. It is crucial to detect the heart disease in patients as soon as possible to prevent heart disease. This research managed to identify the top five significant features for heart disease prediction which are thalach, thal, ca, oldpeak and exang.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the University of Malaya for the FRGS research grant (Project No.: FP057-2017A) to support this research study.

REFERENCES

Gulbinat, W. (1997). What is the role of WHO as an intergovernmental organisation In: The coordination of telematics in healthcare. *World Health Organisation. Geneva, Switzerland at http://www. hon. ch/libraray/papers/gulbinat. html*.

Narain, R., Saxena, S., & Goyal, A. K. (2016). Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. *Patient preference and adherence, 10*, 1259.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering, 106*, 212-223.

Rajeswari, K., Vaithiyanathan, V., & Neelakantan, T. (2012). Feature selection in ischemic heart disease identification using feed forward neural networks. *Procedia Engineering, 41*, 1818-1823.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications, 40*(10), 4146-4153.

Latest Techniques on Entity Detection in Opinion Mining: A Review

Nurul Iva Natasha Bt Moin and Kasturi Dewi Varathan
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya

Corresponding Emails: ivanatasha811@gmail.com, kasturi@um.edu.my

Keywords: *Entity detection, Opinion Mining*

INTRODUCTION

Opinion mining, which is known as sentiment mining, is the science of text analysis for a deeper understanding of the review that was written behind public opinion. The opinion target is usually a named entity such as an organization, individual or event. For instance, in the sentence "The photos quality of my new Sony camera is excellent" the entity for which an opinion is expressed is the Sony camera, the aspect of the entity is the photos quality and the polarity of the opinion is positive. This paper provides a review of the entity detection in opinion mining. This review covered the latest techniques that are used in the past three years (2106-2018) for entity detection in opinion mining. The survey focuses on English language only. Omitting the studies made in another language other than English and entity detection that was obtained in general text such as web documents, news or any other scientific text as it is not within the scope of this study. The need of identifying the right entity in opinion mining is to know what exactly the opinion holders' are talking about in the opinion sentences.

ENTITY DETECTION IN OPINION MINING TECHNIQUES

In this section, the techniques used in opinion mining for entity detection in past three years (2015-2018) are presented. Different techniques have been employed to address the task of entity detection in opinion mining. The following content focuses on the recent techniques that are used by the past researchers.

CONDITIONAL RANDOM FIELDS (CRF)

The Conditional Random Fields (CRF) is a class of statistical modelling method often applied in pattern recognition and machine learning and used for structured prediction. It is used to encode known relationships between observations and construct consistent interpretations. Specifically, CRF find applications in Part-of-Speech (POS) Tagging, shallow parsing, named entity recognition, gene finding and peptide critical functional region finding, among other tasks, being an alternative to the related hidden Markov models (HMMs).

CRF were used by Xu et al.(2016); Xu et al.(2017) in identifying entities in opinion mining. Both research uses 7 popular products with about 1200 reviews that was frequently mentioned from the Amazon review datasets.

In research (Xu, Xie, Shu, & Yu, 2016), it addressed the Complementary Entity Recognition (CER) because recognizing complementary entities is an important task in text mining. The techniques used are CRF and CER6K+ to obtain the most accurate result in entity detection. The acronym for CER6K+ is Complementary Entity Recognition with expanding domain knowledge of 6K reviews and samples of candidate complementary entities and domain-specific verbs. The accuracy for CRF with F1- score is 0.55, CRF perform relatively good on these products, but the performance drop for the last 3 products because of the domain adaptation problem. For CER6K+, the accuracy of F1-score is 0.78. Thus, it shows that CER6K+ performs well on all products.

For the research of (Xu, Shu, & Yu, 2017), it address the Complementary Entity Recognition (CER) as a supervised sequence labelling with the capability of expanding domain knowledge as key-value pairs from unlabelled reviews, by automatically learning and enhancing knowledge-based features. The techniques used are CRF as the base learner, and

extension for CRF with knowledge-based features is KCRF. The accuracy for CRF with F1-score is 0.62 and for KCRF the F1-score is 0.77. KCRF performs well on F1-score because it can automatically identify knowledge-based features and expand knowledge as key-value pairs from plenty of unlabelled reviews. In addition, one of the positive effect for KCRF is that the expanded knowledge is useful in improving the performance of predictions, especially for products without training data.

The reason why CRF have lower accuracy than KCRF and CER6K+ is because CRF is base learner especially when adding new data to the training dataset, It force to re-train the whole CRF model and it may be quite time-consuming due to the high complexity of the training phase of the algorithm.

TARGET IDENTIFICATION BIDIRECTIONAL GATED RECURRENT UNIT (TI-BIGRU)

Target identification aim to extract the targets which customers expressed their opinions on. TI-biGRU have used a bidirectional gated recurrent neural network to extract the targets from a given text. The effectiveness of TI- biGRU is used to automatically identify the target from a tweet.

Ti-biGRU were used in (Jabreel, Hassan, & Moreno, 2018) for target identification. This research uses a publicly available IMDB dataset of extracted adverb-adjective pairs from movie reviews. The dataset contains 6248 training examples and 692 examples in the testing set. The techniques used is Target Identification-Gated Recurrent Unit (TI-GRU) and TI-biGRU. TI-GRU is the simplified version of TI-biGRU in which only the past information is considered, ignoring the directionality. Scores for TI-biGRU's precision, recall and F1-measure are 87.39, 91.18 and 89.25 respectively. Likewise, the scores for TI-GRU's precision, recall and F1-measure are 81.8, 90.89 and 89.25 respectively. It is clearly shown that TI- biGRU outperforms the other models.

LIFELONG POSITIVE-UNLABELLED (LPU)

LPU algorithm consists of four main steps: knowledge accumulation, current domain setup, knowledge mining and preparation, and restricted PU iterations. LPU were used in (Wang, Zhou, Mazumder, Liu, & Chang, 2018) with two-stage approach which are stage one: extracts/groups the target-related words (call t-words) for a given target. Next is, stage two: disentangling or separates the aspect and opinion words from the grouped t-words, which is challenging because usually do not have enough word-level aspect and opinion labels.

LPU is used in stage two of this research. Dataset used is a large corpus of Amazon reviews from 20 different domains. Specifically, three domains from different product categories are selected, namely, cell phone, beauty and office. The techniques used is LPU and LPU-. Lifelong Positive-Unlabelled minor (LPU-) is a LPU variant that does not make risky self-prediction exploration and relies more on the past mined knowledge. The result for the accuracy in LPU and LPU- outperform other baselines markedly. LPU results by 0.83, 0.86 and 0.90 in acc@150, acc@100, and acc@50 respectively. Meanwhile, LPU- results by 0.80, 0.85 and 0.94 in acc@150, acc@100, and acc@50 respectively.

Table 1: The Accuracy of the Techniques

Paper	**Techniques**	**Accuracy (%)**	**Domain**
[4]	CRF	55	Gadget Accessories
	CER6K+	78	
[3]	CRF	62	
	KCRF	77	

[1]	TI-GRU	89.25	Movie review
	TI-biGRU	89.25	
[2]	LPU	90	Cellphone, Beauty, and office
	LPU-	94	

ENTITY DETECTION ANALYSIS

The highest accuracy is Learning Positive Unlabelled minor (LPU-) with 94% and the lowest is Conditional Recurrent Field (CRF) with 55%. The domain is based on product and reviews. Product is catered into gadget, beauty and office. Generally, gadget based product is used in this reviewed research.

CONCLUSION

Research on entity detection has less been discovered. It is an important area that needs to be looked into. This is because entity detection in opinion mining is vital in this area in which to know what exactly the opinion holder is talking about in their opinions. Many research works are found using their own dataset that is not publicly available. Different dataset of each research made it difficult to the new researcher to determine which technique to be used, as the accuracy may be differ using different dataset. On top of that, entity detection needs to be researched further as it contains the least amount of research works.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the University of Malaya Research Grant (Project No.: RP059D-17SBS) to support this research study.

REFERENCES

Jabreel, M., Hassan, F., & Moreno, A. (2018). Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. In *Smart Innovation, Systems and Technologies* (Vol. 85, pp. 39–55). https://doi.org/10.1007/978-3-319-66790-4_3

Wang, S., Zhou, M., Mazumder, S., Liu, B., & Chang, Y. (2018). Disentangling Aspect and Opinion Words in Target-based Sentiment Analysis using Lifelong Learning. Retrieved from http://arxiv.org/abs/1802.05818

Xu, H., Shu, L., & Yu, P. S. (2017). Supervised Complementary Entity Recognition with Augmented Key-value Pairs of Knowledge. Retrieved from http://arxiv.org/abs/1705.10030

Xu, H., Xie, S., Shu, L., & Yu, P. S. (2016). CER: Complementary Entity Recognition via Knowledge Expansion on Large Unlabeled Product Reviews.

Spatial Big Data for Coastal Erosion Mitigation and Prediction

Patrice Boursier, Raja Kumar Murugesan, Venantius Kumar Sevamalai, Lim Eng Lye, Sohaib Al-Yadumi and Denis Delaval
Taylor's University, Malaysia

Corresponding Email: patrice.boursier@taylors.edu.my

With sea level rising, the number of people living in coastal areas is continuously and rapidly growing. Combined with storms, floods and erosion, this has a strong impact on coastal populations, infrastructures and ecosystems. Strategies and solutions depend on the availability of datasets related to the topography, the meteorology and sea data (sea level, waves, currents, …). These data need to be combined and analysed effectively in order to derive knowledge and prediction.

Big Data and Internet of Things (IoT) technologies (devices and software) can help attain the objectives. Big Data relates to the storage, processing and analysis of large volumes and varieties of data, in a fast and reliable manner. IoT relates to the real-time collection of data using sensors connected to servers. Spatial Big Data solutions combined with IoT will help predict and mitigate risks related to floods and coastal erosion.

Working on coastal erosion and floods implies that we work with geo-referenced or so called spatial data. They refer to digital maps, satellite images and geo-referenced data sets in general (different kinds of statistical or descriptive data, related for example to population or socio-economic data, the quality of water or the air, wind measurements, height of the waves, etc). This means high volumes of structured, semi-structured and unstructured data. These data are heterogeneous in type, format and quality.

Work has been done on the integration of spatial data sets, but there are still many issues that have to be solved, mostly related to high volumes, data quality, data integration, data analysis and data visualization. It is therefore necessary to develop specific methodologies for collecting (IoT), checking the quality and correcting (big data cleansing), integrating (big data management), analysing (big data analytics) and visualizing (3D, animated visualization) these data sets.

These are common big data problems, but we need to develop more advanced analytical tools for spatial data sets that propose unique challenges. Therefore, we need an integrated spatial-social-network model. Spatial and social network models typically do not operate on the same scale, and they do not make consistent predictions. Moreover, current analytical models do not scale to the size of current data sets.

The overall project methodology consists in developing in parallel methodologies for (i) analysing and visualizing environmental data sets for coastal erosion, (ii) integrating spatial big data sets, and (iii) developing a spatial IoT allowing to collect in real time data related to coastal erosion. Then, a prototype will be implemented in order to validate the approach and methodologies.

REFERENCES

Alexander G. Rumson, Stephen H. Hallett, Timothy R. Brewer (2017). Coastal risk adaptation: the potential role of accessible geospatial Big Data. Marine Policy, 83, pp. 100–110, Elsevier.

Yang, M. Yu, F. Hu, Y. Jiang, Y. Li (2017). Utilizing cloud computing to address big geospatial data challenges. Computers Environment and Urban Systems, 61, pp. 120–128, Elsevier.

J. Georis-Creuseveau, C. Claramunt, F. Gourmelon (2017). A modelling framework for the study of spatial data infrastructures applied to coastal management and planning. Int. Journal of Geographical Information Science, 31, pp. 122 138, Taylor and Francis.

Huw Vaughan Thomas (2016). Coastal Flood and Erosion Risk Management in Wales. Wales Audit Office, July 2016.

H. de Moel, B. Jongman, H. Kreibich, B. Merz, E. Penning-Rowsell, P.J. Ward (2015). Flood risk assessments at different spatial scales, Mitigation and Adaptation Strategiescfor Global Change, 20, pp. 865–890, Springer.

Kuien Liu, Yandong Yao, Danhuai Guo (2015). On managing geospatial big-data in emergency management: some perspectives. 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management (EM-GIS '15), Washington, November 2015.

J.-G. Lee, M. Kang (2015). Geospatial Big Data: challenges and opportunities. Big Data Research, 2, pp. 74–81.

R. Devillers, D.M. De Freitas (2013). The use of GIS and geospatial technologies in support of coastal zones management – results of an international survey, 11th International Symposium on GIS Computing and Cartography for Coastal Zone Management, CoastGIS, Victoria, British Columbia, Canada, pp. 100–103.
Elisabetta Genovese, Valentin Przyluski (2012). Storm surge disaster risk management: the Xynthia case study in France. Journal of Risk Research, pp. 1–17, Taylors and Francis.
Xin Chen, Hoang Vo, Ablimit Aji, Fusheng Wang (2014). "High Performance Integrated Spatial Big Data Analytics". 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial).
V. Bhanumurthy, K. Ram Mohan Rao, G. Jai Sankar, P. V. Nagamani. "Spatial data integration for disaster/emergency management: an Indian experience". Spatial Information Research. April 2017, Volume 25, Issue 2, pp 303–314.
Nicolas Becu, Marion Amalric, Brice Anselme, Elise Beck, Xavier Bertin, Etienne Delay, Nathalie Long, Nicolas Marilleau, Cécilia Pignon-Mussaud and Frédéric Rousseaux (2017). "LittoSIM: a participatory simulation to foster social learning on coastal flooding prevention".
Michael F. Goodchild (2016). "GIS in the Era of Big Data". Cybergeo : European Journal of Geography. Published online 25 April 2016, read 23 May 2017.
T.W. Gallien (2016). Validated coastal flood modelling at Imperial Beach, California: Comparing total water level, empirical and numerical overtopping methodologies. Coastal Engineering, 111, Elsevier, 95–104.
Zaharah Eliasa, Zaiton Hamin, Mohd Bahrin Othman (2013). "Sustainable Management of Flood Risks in Malaysia: Some lessons from the legislation in England and Wales". Procedia - Social and Behavioral Sciences, 105, Elsevier, 491 – 497.
Huw Vaughan Thomas (2016). "Coastal Flood and Erosion Risk Management in Wales". Wales Audit Office.

The Impact of Dominant Color in Online Advertising on Purchase Intention: A Preliminary Study

Fatemeh Bakhshian and Wai Lam Hoo
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya

Corresponding Emails: fbakhshian66@gmail.com, wlhoo@um.edu.my

INTRODUCTION

In recent years, online advertising revolutionized the marketing by creating a opportunities for advertisers to reach potential customers. It is intuitive that "where the eye stops, the sale begin" (Kauppinen-Räisänen, 2014)Therefore, advertising draw potential customers' attention in order to influence them to buy a product (Estrada-Jiménez, Parra-Arnau, Rodríguez-Hoyos, & Forné, 2017).Advertisers have long been interested in ways to attract attention. The choice of colors to include in an advertisement is an important issue advertising practitioners, since color is highly visible element in every form of visual communication. Online purchase intention as a key requirement of a transaction, is a consumer's desire to buy a product or service from a web site. (Shaouf, Lü, & Li, 2016). When consumers buy goods, they search for relevant information based on their experience and the external environment. After obtaining a certain amount of relevant information, consumers begin to evaluate and consider the product and, after making comparisons and judgments, will engage in purchase behavior. Defined purchase intention as a transaction behavior consumers tend to exhibit after evaluating a product, and adopted consumer reaction to a product to measure consumer purchase likelihood. Higher purchase intention means higher likelihood of consumers purchasing a product (Wang, Cheng, & Chu, 2013).

PROBLEM STATEMENT

Many research studied the importance of color in marketing and advertising. In designing ads, one of the decisions the advertiser must make is which color(s) to use as executional cues in the ad. To provide guidelines for these decisions , Gorn et. al (1977), proposes and tests a conceptual framework linking the hue, chrome, and value of the color(s) in an ad to consumers' feelings and attitudes. The three key ad elements (brand, pictorial, and text) each have unique effects on attention to advertisements, are examined by Pieters, et al,(2004), in which shows The pictures capturing more attention, independent of its size . The impact of color on associations evoked by different types of images evaluated by (Kuzinas, 2013) . Results revealed that red and blue colors evoked different associations and this difference remained despite the effects of other image elements. According to Hsieh, et al (2018) online consumers' reactions to online merchandise prices vary according to website background color. Sliburyte et al,(2014) explore the consumer color perception. Results presented in this paper revealed that color perception only partially depends on demographic factors, because color is perceived differently by people of different age, gender and education. However, the impact of specific dominant color used in online advertisements on potential customers purchase intention have not investigated in the literature.

LITERATURE REVIEW

Advertising: According to Liu et al, (2018) online advertising is defined as any paid form of information about products in an online environment without geographical boundary limits . In an-other word, it is a form of promotion that uses the Internet and World Wide Web for the express purpose of delivering marketing messages to attract customers. (Kim, Kwon et al.2011) Advertising linked to marketing-specific branding by drawing attention of potential customers in order to influence users to buy a product and generally spreading (Estrada-Jiménez et al., 2017).

Color: By definition, color is that part of perception that is carried to us from our surroundings by differences in the wavelengths of light, is perceived by the eye, and is interpreted by the brain (Panigyrakis & Kyrousi, 2015). Consumers have different psycho-logical characteristics in the different stages of the process of accepting advertising, that guide Internet users to speed up their attention, interest, association, desire, motivation and purchase decision (Wu, 2017).

Purchase intention: It indicates likelihood that consumers will plan or be willing to purchase a certain product or service in the future (Martins, Costa, Oliveira, Gonçalves, & Branco, 2018). There are numerous subject and attribute in online

marketing and advertisement which affect purchase intention, including: motivational factors (Kim, Kim, & Park, 2010), techno-product (Chen, Hsu, & Lin, 2010), and Facebook advertising.(Dehghani & Tumer, 2015)

METHODOLOGY

Based on the literature, a set of criteria that affects the pur-chase intention is identified. Then, an expert system or fuzzy interference system is designed to 1) identify the dominant color of the online advertisement; 2) identify the type of line advertisement; and 3) estimate the purchase intention of the customer. The online advertisement is served as the input as digital image and the output would be the dominant color identified and the purchase intention of the customer. Relevant technical comparison will be identified.

FINDINGS

18 criteria that are identified which are related to dominant color identification, online advertising type, and customer purchase intention. A set of rules for expert system (or fuzzy rules for fuzzy inference system) will be developed based on these criteria.

VALUE

This project proposed an automated system that identifies the effects of color on customer purchase intention, within online advertising domain. In specific, dominant color identification is proposed because it is believed that the effectiveness of on-line advertising are strongly dependent on the choice of color that represent the advertisement as a whole.

REFRENCE

Chen, Y.-H., Hsu, I.-C., & Lin, C.-C. (2010). Website attributes that increase consumer purchase intention: A conjoint analysis. *Journal of Business Research, 63*(9-10), 1007-1014.

Dehghani, M., & Tumer, M. (2015). A research on effectiveness of Facebook advertising on enhancing purchase intention of consumers. *Computers in Human Behavior, 49*, 597-600.

Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A., & Forné, J. (2017). Online advertising: Analysis of privacy threats and protection approaches. *Computer Communications, 100*, 32-51.

Kauppinen-Räisänen, H. (2014). Strategic use of colour in brand packaging. *Packaging Technology and Science, 27*(8), 663-676.

Kim, J. U., Kim, W. J., & Park, S. C. (2010). Consumer perceptions on web advertisements and motivation factors to purchase in the online shopping. *Computers in Human Behavior, 26*(5), 1208-1222.

Martins, J., Costa, C., Oliveira, T., Gonçalves, R., & Branco, F. (2018). How smartphone advertising influences consumers' purchase intention. *Journal of Business Research*.

Panigyrakis, G. G., & Kyrousi, A. G. (2015). Color effects in print advertising: a research update (1985-2012). *Corporate Communications: An International Journal, 20*(3), 233-255.

Shaouf, A., Lü, K., & Li, X. (2016). The effect of web advertising visual design on online purchase intention: An examination across gender. *Computers in Human Behavior, 60*, 622-634.

Sliburyte, L., & Skeryte, I. (2014). What we know about consumers' color perception. *Procedia-Social and Behavioral Sciences, 156*, 468-472.

Wang, J. S., Cheng, Y. F., & Chu, Y. L. (2013). Effect of celebrity endorsements on consumer purchase intentions: advertising effect and advertising appeal as mediators. *Human Factors and Ergonomics in Manufacturing & Service Industries, 23*(5), 357-367.

Wu, Y. (2017). 85. Design and Implementation of Virtual Advertising Based on Visual Communication Design. *Boletín Técnico, ISSN: 0376-723X, 55*(18).

The Impact of Machine Learning on Economics

Maryam Moradbeigi[1] and Mohsen Saghafi[2]
[1]Taylor's Business School, Taylor's University, 47500 Subang Jaya, Selangor, MALAYSIA
[2]Tech Lead, Supahands Sdn. Bhd.

Corresponding Email: Maryam.Moradbeigi@taylors.edu.my

INTRODUCTION

The ongoing impact of Machine Learning (ML) on variety of economic topics has been widely recognized. This short report aims to give some insights of the use of ML in economics. Then, it follows by some problems facing the implication of ML in this field of study.

MACHINE LEARNING IN ECONOMICS

The designed algorithms in machine learning are mainly utilized in clustering or classifying a dataset or applied in prediction. Generally, the main focus of machine learning can be classified under two groups, e.g. supervised and unsupervised machine learning.

Unsupervised machine learning includes variety of techniques such as k-means clustering and topic modelling to classify the dataset into different clusters for each of them the covariances of observations are similar. (Blei et al., 2003) This is a helpful tool in economic analysis to construct both dependent and independent variables. As an example, if a company is interested in the factors affecting the customers' shopping behaviour, the data-driven unsupervised ML technique can be used to cluster the items with similar characteristics from an online review. Thus, it will be possible to find out what kind of features make customers to view products as a high quality ones. The advantage of this method is the researcher is able to avoid human judgment into the analysis (Athey et al., 2017)

Predicting an endogenous variable (Y) using the explanatory variables (X) is a typical purpose of adopting supervised ML in economics (Mullainathan and Spiess, 2017; Varian, 2014; White, 1992). To this end, the machine uses the training dataset to set the best coefficients. Then, the accuracy of the system is measured by running the algorithm on test dataset. Therefore, it is noteworthy to mention that the estimation (i.e. measuring changes in when X changes by one unit) is not a concern in this process, but rather the goodness of fit, that is the minimum differences between actual and predicted dependent variables in independent test set, is the main goal of supervised ML.

TRADITIONAL ECONOMETRICS AND MACHINE LEARNING

The ML techniques contrast the traditional econometrics in two main ways.

The causal inference is the primary goal in applied econometrics. In other words, the applied econometrics concerns with how changing one independent variable impacts the dependent variable while other variables are kept constant. As an instance, the Romer growth model theoretically shows the technology advancement positively affects the pace of growth. A researcher may empirically investigate how big the impact on economic growth would be if the expenditure on R&D changes. However, the explicit goal of machine learning is goodness of fit by using flexible functional form (Athey et al., 2016; Zubizarreta, 2015).

This leads us to the second difference between ML and traditional econometric. A researcher creates a model in empirical econometrics analysis based on the theories. For example, the Romer model proposed the factors associating with economic growth. So, the researcher constructs the model based on this theory and is not allowed to change the model until they get the favourable results. Then, they will use the available data to estimate the model

and statistical theories play crucial role in finding the confidence intervals, analysing the potential biased estimation and test the reliability of estimated parameters. They could, however, report two or three other models as sensitivity analysis. However, the algorithms in ML is used to estimate variety of alternative models and compare the result to find the best fit and it does not bother with the uncertain estimated coefficients. Thus, one may conclude that traditional econometric is theory driven while ML is data driven methodologies.

CONCLUSION

The impact of machine learning on economics has become profound. Applying ML in the empirical economic researches is very vibrant. And it can be expanded in dozens of other emerging topics. In conclusion, I believe that fundamental transformation in applied econometrics.

REFERENCE

Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*,2016.

Athey, M. M. Mobius, and J. Pál. The impact of aggregators on internet news consumption. 2017.

M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

S. Mullainathan and J. Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

H. R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28 (2):3–27, 2014

H. White. *Artificial neural networks: approximation and learning theory*. Blackwell Publishers, Inc., 1992

R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. doi: 10.1080/ 01621459.2015.1023805.

The Socio-Technical for Cyber Propagation in Social Media: An Integrative Model of Human Behavior and Social Media Power in Cyber Propagation

Aimi Nadrah Maseri and Azah Anir Norman
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Email: azahnorman@um.edu.my

Keywords: *Propaganda, socio-technical, social media, behavior.*

INTRODUCTION

The cyber information propagation behavior and sophisticated social media power in the world history and public memory has come to the fore in recent years, especially with the rise of multi-capabilities of social networking tools such Tweetdeck by Twitter and LikeAnalyzer for Facebook. Cyber propaganda is asserted to become a highly potential security threat and could be more damaging to the global stability (Bharat, 2017). Howard et. al (2011), comparative analysis shows both democratic and authoritarian regimes disable social media due to security concern, with cyber-propaganda as one of the citing concern There are many other series of cyber-propaganda events where the biggest threat is in influencing Politics 2016 (Hacquebord, 2017). Due to these many reasons, it is significant for the research to investigate the cyber information propagation behavior and sophisticated social media power, hence understand the magnifying ripple effect associated with this matter.

PROBLEM STATEMENT

The research embarks on the objectives to investigate the interaction and association between cyber information propagation behavior and social media power, to determine the information propagation ripple effect and to identify cyberpropaganda determinants.

METHODOLOGY

This study employ a mixed methods sequential design which combining both qualitative and quantitative research methodologies.

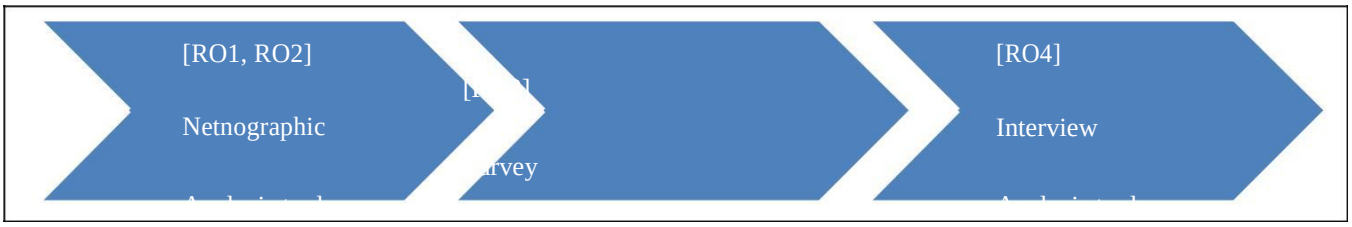


Figure 1: Research Design

The first phase will be involved netnographic method in order to observe the user behavior in social media towards propaganda messages. Netnographic method is applied to understand social interaction in the Internet context (Kozinets, 2015). To perform this methodology, a Facebook group that is likely to spread propaganda will be chose, and the social network analysis (SNA) will be conducted using nodeXL to identify and categorize opinions expressed in a comment, in order to determine whether the user's attitude towards a particular topic is

positive, negative, or neutral. Then, a survey will be conducted in order to proposed a socio-technical model in cyber propagation context. The questionnaire consisted of an introductory letter, a concise description of propaganda in social media, queries about demographic information, basic habits of using social media and measures for the research variables. The initial version of the questionnaire was tested for content validity by three PhD candidates and two post-graduate students not involved in this research, and a minor amendment was made according to their feedback. The last phase is the validation of the model. The interview will be conducted and interviewee will be identified and selected based on the expertise in propaganda and social media. The interview protocol also will be created to assist and to ensure the objective of the interview will be achieved.

EXPECTED FINDINGS

The extant research on the formation of online social networks is mainly pursued from network-level perspectives. Our research contributes to the society through knowledge transfer on mitigating threats in cyber-propaganda activities by providing cyber-propagation information analysis metrics/tool, which is the ripple effect model that associates the cyber information propagation behaviour and social media power. Cyber-propaganda may influence a nation through misinformation, therefore through this research, a metric is proposed to analyse the level of severity of specific propaganda, hence economically save many unintended resources, due to the impact of the aftermath of misinformation. Our research can also provide useful insights for practitioners. With a better understanding of the key factors that influence cyber-propaganda, users, organizations or government will be able to better educate and inspire others to combat misinformation. Considering the positive impacts of awareness of adverse consequences, social media users can be directed to spread accurate information, by improving the awareness of the side effects caused by propaganda.

Utilizing the Data Socialization for Predicting Future Trends in Social Entrepreneurship

Nur Azreen Zulkefly and Norjihan Abdul Ghani
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Email: norjihan@um.edu.my

Keywords: *Social Entrepreneurship, Predicting Future Trends, Data Socialization*

INTRODUCTION

According to UN Global Pulse (May 2012), Big data analytics refers to tools and methodologies that aim to transform massive quantities of raw data into "data about the data" for analytical purposes. In the field of social entrepreneurship, data analytics can assist entrepreneurs in predicting future trends. It can help social entrepreneurs make informed decisions to drive the company forward, improve efficiency, increase profits and achieve organisational goals and also helps society concurrently, Margaret Rouse (December, 2016). The current problem in social entrepreneurship is they do not use social impact data effectively when making decisions, Kate Ruff, (May 2016). With data socialization, business users and data analysts can be more productive and better connected as they source, cleanse and prepare data for analytical and operational processes, Jon Pilkington, (2017). Predicting future trends is based on the idea that what has happened in the past gives traders an idea of what will happen in the future, Faraz Rasheed et al. (2014). Predicting future trends in social entrepreneurship can apply data socialization approach such as spotting and monitoring behaviours and patterns allows us to take a stab at predicting where things are heading, how demand for our products or services will change over time, and what will prompt that change.

PROBLEM STATEMENT

Predicting future trends in social entrepreneurship is important as it promote an in-depth analysis related to the history of social entrepreneurship in order to determine what is going wrong and to discover new ways to solve social problems. Predicting future trends using data socialization approach is yet available in social entrepreneurship domain. Therefore, using this approach can address social entrepreneurs in assessing the social impact of their business in the future because this approach can make better decisions about which data to use in analytics processes.

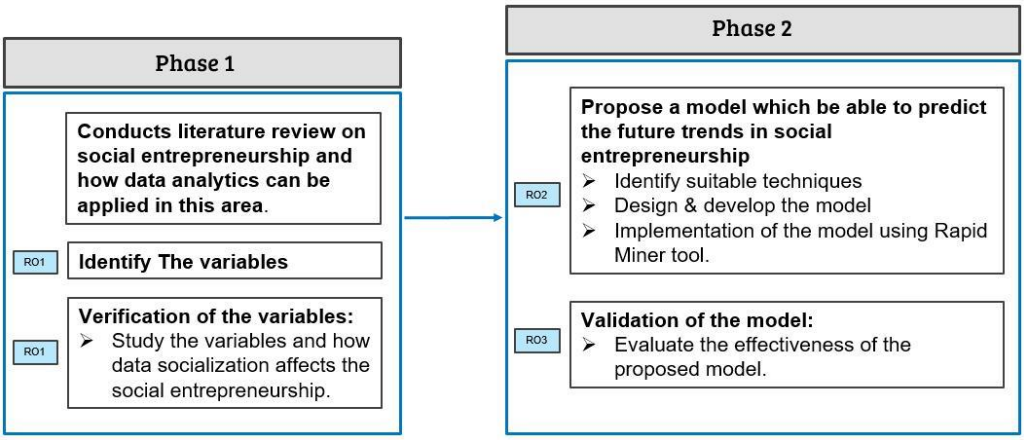


Figure 1: Research Design

METHODOLOGY

The research design of this research are divided into two major phases. The first phase involve three main tasks which are literature study on how data analytics can be applied in social entrepreneurship domain, identification of the variables and verification of the variables which enable to predict future trends in social entrepreneurship. These task will be conducted by study on the variables involve in social entrepreneurship and how it is going to affect the future trends. The variables will be observed based on the existing variables used by Small Medium Enterprise (SME) and Small Medium Industry (SMI). The research then continue with the second phase which include two major tasks. The tasks are include to propose a model that can predict future trends in social entrepreneurship. The model is going to be implement using Rapid Miner

tool. The last task is the validation of the model. This task is going to be conducted to evaluate the effectiveness of the proposed model.

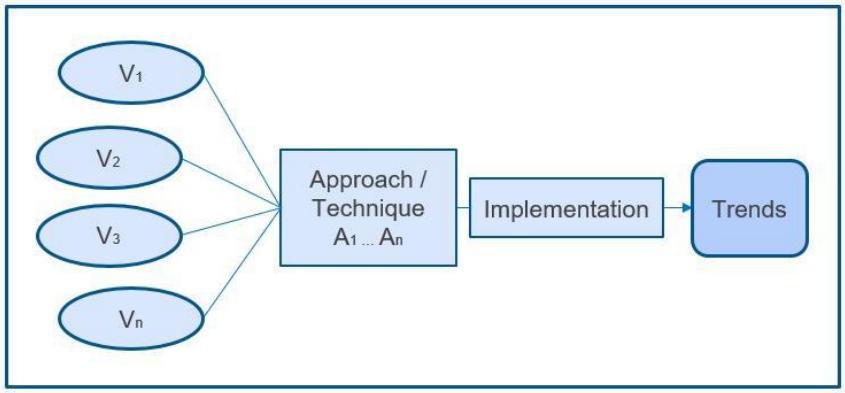


Figure 2 Conceptual Model

FINDINGS

The expected findings of this research is to propose a model which can be used by social entrepreneurs in predicting the future trends of their business. At the end of this research, social entrepreneurs can applied the data socialization proposed through the model in assisting the entrepreneurs. Based on Figure 2, the diagram shows the conceptual model of this research. The conceptual model is the representation of the model that will be develop throughout this research. The first phase of the model is to identify and verify the variables in social entrepreneurship. Then the variables will apply suitable approach/ techniques and then it will be implement using Rapid Miner tool. The results of this model is the future trends that can be used by social entrepreneur to solve social problem in their business.

A Unified Model of STEM Game Based Learning Apps to Enhance Creativity among Preschoolers

Najmeh Behnamnia
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Emails: n.behnamnia@gmail.com, amir@um.edu.my

INTRODUCTION

Recently, the trend of using game based learning apps through touch screen devices is among young children has been increasing. Most of preschoolers have access to a device, such as tablet and smartphone (touch screen) at home or school. These devices (tablet and smartphone) have become essential part of many young children daily routine [5]. According to the survey of 1028 children (3-5) years was carried out by the National Literacy Trust demonstrated that more than 70% have access to a device with a tablet and smartphone (touch screen) at home and school [6]. In addition, the number of games based learning apps in the market is increasing. The attention of parents and teachers to this kind of apps (educational game) is rising too. Schuler 2012 reported that more than 80% of the top-selling apps in the Apple Apps Store in pre-school, education has been targeted. So, to see how these apps have been selected and used, research and analysis is needed [7, 8]. In addition, teachers and parents have repeatedly asked researchers to investigate the effects of educational programs among young children. Due to the large number of these requests, an essential demand for study on the use of media and technology in preschool age were observed [9, 10 & 11]. A wide range of prior studies on the use educational games for Pre-school, have largely focused on the training in the basic settings and not expressly on game analysis and creativity [12, 13& 14]. Also, due to the increasing of using of technology among children, play and foster creativity in the virtual environment has become a challenge to investigate [14, 15 & 16]. Few studies have pointed out that young children can use an extensive area of technologies in order to elevate creativity [17]. However, further research on the variety, creativity and nurture it during games training is still needed [17 & 18].

PROBLEM STATEMENT

The gabs identified are based on the review of previous studies that can be divided into several categories as follows:

To date, there has been little agreement on what bounds the definition of an educational game or play in preschooler's level [19]; scholars have used different definitions for the words play and game. Rarely are game- and instructional-design both included in creating educational games [20]. With insufficient emphasis on the context of game-based learning and fostering creativity [21], the field needs a game that has undergone a rigorous research and development (R&D) process from the beginning to the end.

Further, the application of game-based learning and creativity has not been adequately realized among young children in learning environments. This has been in part due to a lack of empirical evidence supporting beneficial claims [22, 23]. Without research providing stakeholders empirically validated evidence of the profound benefits of educational gaming, researcher may continue to miss out on the real solutions play offers 21st century science instruction [24, 25].

Combining both the components of creativity and learning within a single model or framework could provide better performance for preschooler's level [26, 27].

In addition, the existing GBL apps have not successfully match between their design features of fostering creativity, learning, and existing assessing pedagogy for preschooler [4]. Therefore, designing a pedagogy scheme as a facilitator is essential to provide a comprehensive analysis of proposed model.

METHODOLOGY

Participation: The participants recruited in this research project consisted of Montessori preschool. Montessori education for preschooler is fundamentally a method of learning through playing game.

Procedure: The procedure is Case study. Through case study methods, a researcher is able to go beyond the quantitative statistical results and understand the behavioral conditions through the actor's perspective. By including both quantitative and qualitative data, case study helps explain both the process and outcome of a phenomenon through complete observation, reconstruction and analysis of the cases under investigation [88].

Observations & interviews: This step will be through of 'Go Pro' chestcam. Go Pro' chestcam is a camera that is strapped to the child's chest and allows the recording of action as the child moves and interacts with other people and objects, including tablets without their care. A trained researcher will conduct the experiment with each individual child participant in a quiet location. As a warm up, the researcher will engage the child in a few activities to assess the child's verbal ability. Parents of participating children will have completed an online survey about the child's media habits, general behavior, and family demographics. All testing sessions will be video and audio recorded with "chestcam".

Instrument to use apps: in this project we are using a Touchscreen tablet (iPad). The use of traditional computers limits the creation of educational activities targeted to preschool children (since they are placed in a fixed location). Traditional computers use a mediated interaction (mouse and keyboard) which is not very natural or intuitive for this specific type of users who have not fully developed their fine motor skills yet. The use of tabletop technologies solves several limitations of the traditional computers. Changes to a direct-touch approach (tangible or tactile interaction), which is preferred by children. The tabletop's form factor and size offers more opportunities for supporting collaboration between peers. These devices allow the movement of children with the device along the area where the activity is being carried out [28].

Measurement tools:

a. IBM SPSS 22 statistical package: The interview data will be transcribed and imported into Nvivo10 [29]. Post-test "Cramer's V" effect sizes will be calculated [91 &92] for outcome of impact of creativity and learning.

b. Hughes' taxonomy [30]: To classify and exploring play behaviors

c. ACCT Framework [31]: To classify and exploring creative thinking

d. The Picture Naming Individual Growth and Development Indicator [32]: Will be used as a measure of verbal ability.

FINDINGS

Regarding to assess and Foster creativity and STEM Learning our Unified Model of STEM Game Based Learning Apps to Enhance Creativity among Preschoolers (Pedagogy and STEM) the result can be categorized in these section as follow as bellow:

1. Pedagogy Results:

In this section we have two separate assessments that included; 1) Creative teaching and 2) Creative learning that researcher explained two separate assessments as follow as bellow:

1) Creative teaching:

Good: T, Satisfaction/ STE, Attention/ SM, Relevance/ M, Confidence

Poor: M, Attention

2) Creative learning

Good: TEM, Personification / Inspiration, T /Gamification, ES

Poor: Personification, E/ Gamification, M

2. Creativity Results:

In this section again we have two separate assessments that included; 1) Creativity and 2) Play that researcher explained two separate assessments as follow as bellow:

1) Creativity:

Good: E1, T/ E2, S/ E3, E

Poor: E3, M

3) Play:

Good: Exploratory Play, EM/Objective, E

3. Overall Design Game Scores:

Good: Swipe the Screen, Trace shapes with fingers

Poor: Pinching and Dragging, Exit apps and enter

4. Identify Weakness: Curriculum designers incorporates learning materials and activities emphasize relevant aspects of creativity skill development.

REFERENCES

1. Prensky, M. and M. Prensky, Digital game-based learning. Vol. 1. 2007: Paragon house St. Paul, MN.
2. Wilson, A., T. Hainey, and T.M. Connolly, Using Scratch with primary school children: an evaluation of games constructed to gauge understanding of programming concepts. International Journal of Game-Based Learning (IJGBL), 2013. 3(1): p. 93-109.
3. Rideout, V., Learning at home: Families' educational media use in America. Joan Ganz Cooney Center, 2014.
4. Fisch, S.M., Children's learning from educational television: Sesame Street and beyond. 2014: Routledge.
5. Anderson, D., et al., J. C.(2001). Early Childhood Television Viewing and Adolescent Behavior: The Recontact Study. Monographs of the society for Research in Child Development. 66(1): p. 264.
6. Dingwall, R. and M. Aldridge, Television wildlife programming as a source of popular scientific information: A case study of evolution. Public understanding of Science, 2006. 15(2): p. 131-152.
7. Fisch, S.M. and S.K. McCann, Making broadcast television participative: Eliciting mathematical behavior throughSquare One TV. Educational Technology Research and Development, 1993. 41(3): p. 103-109.
8. Linebarger, D.L., et al., Effects of Viewing the Television Program Between the Lions on the Emergent Literacy Skills of Young Children. Journal of Educational Psychology, 2004. 96(2): p. 297.
9. Mares, M.-L. and E. Woodard, Positive effects of television on children's social interactions: A meta-analysis. Media Psychology, 2005. 7(3): p. 301-322.
10. Jennings, N.A., S.D. Hooker, and D.L. Linebarger, Educational television as mediated literacy environments for preschoolers. Learning, Media and Technology, 2009. 34(3): p. 229-242.
11. Borzekowski, D.L. and J.E. Macha, The role of Kilimani Sesame in the healthy development of Tanzanian preschool children. Journal of Applied Developmental Psychology, 2010. 31(4): p. 298-305.
12. Hays, R.T., The effectiveness of instructional games: A literature review and discussion. 2005, DTIC Document.
13. Squire, K. Video games in education. in International journal of intelligent simulations and gaming. 2003. Citeseer.
14. Villalta, M., et al., Design guidelines for classroom multiplayer presential games (CMPG). Computers & Education, 2011. 57(3): p. 2039-2053.
15. Lester, J.C., et al., Designing game-based learning environments for elementary science education: A narrativecentered learning perspective. Information Sciences, 2014. 264: p. 4-18.
16. Anderson, S.P., Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences, Portable Document. 2011: Pearson Education.
17. Squire, K., Video Games and Learning: Teaching and Participatory Culture in the Digital Age. Technology, Education—Connections (the TEC Series). 2011: ERIC.
18. Flow, C., The psychology of optimal experience. Harper&Row, New York, 1990.
19. Shute, V.J., Stealth assessment in computer-based games to support learning. Computer games and instruction, 2011. 55(2): p. 503-524.
20. Boyle, E., T.M. Connolly, and T. Hainey, The role of psychology in understanding the impact of computer games. Entertainment Computing, 2011. 2(2): p. 69-74.
21. Wu, W.-H., et al., Re-exploring game-assisted learning research: The perspective of learning theoretical bases. Computers & Education, 2012. 59(4): p. 1153-1161.
22. Wu, W.H., et al., Investigating the learning theory foundations of game based learning: a meta-analysis. Journal of Computer Assisted Learning, 2012. 28(3): p. 265279.
23. Li, M.-C. and C.-C. Tsai, Game-based learning in science education: A review of relevant research. Journal of Science Education and Technology, 2013. 22(6): p. 877-898.
24. Vgotsky, L., Mind in society: The development of higher mental processes. 1978, Cambridge, MA: Harvard University Press.
25. Squire, K., Changing the game: What happens when video games enter the classroom. Innovate: Journal of online education, 2005. 1(6).
26. Gee, J.P., What would a state of the art instructional video game look like? Innovate: Journal of online education, 2005. 1(6): p. 1.
27. Twining, P., Exploring the educational potential of virtual worlds—Some reflections from the SPP. British Journal of Educational Technology, 2009. 40(3): p. 496-514.
28. Twining, P., Virtual worlds and education. 2010, Taylor & Francis.
29. Johnson, P. (2009). The 21st century skills movement. Educational Leadership, 67(1), 11–11.
30. Trilling, B., & Fadel, C. (2009). 21st century skills: Learning for life in our times. San Francisco, CA: Jossey-Bass.
31. Thomas, B. W. (2007). Creative cognition as a window on creativity. Methods (San Diego, California), 42(1), 28–37. doi:10.1016/j.ymeth.2006.12.00
32. Lewis, T. (1999). Research in technology education: Some areas of need. Journal of Technology Education, 10(2), 41–56.

Using Regression Models in Calculating Iterative Learning Control (Ilc) Policies for Fed-Batch Fermentation

J. Jewaratnam and J. Zhang
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Corresponding Email: jegalaxmi24@um.edu.my

Keywords: *Fed-batch Fermentation, Iterative Learning Control, Multiple Linear Regression, Partial Least Square, Principal Component Regression*

Most of the bio-products including proteins, biopolymers and primary and secondary metabolites are being produced using fed-batch fermentation. Fed-batch fermentation is a process in which substrates are fed into the system at different intervals. Fed-batch fermentation is generally more efficient compared to batch fermentation. However, it is a dynamic system which is prone to have uncertainties due to nonlinear behaviors of living organisms.

Conventional control system is not able to tackle the model dynamics of the fed-batch system. Due to limited availability of the robust on-line sensors for fed-batch system, offline measurements are the key indicators of the product quality. This is the common method for direct measurement of variables such as feed rate, temperature and concentration for tracking purpose. As the desired output is fixed, the same time-varying trajectories are used batch after batch to achieve desired output. This approach works well for the system in which the disturbances affect the controlled process variables first, followed by the product quality through the kinetics. The qualities of products are usually affected by more than one factor. This means that maintaining the measured variables alone does not ensure desired final product quality.

The practical difficulties faced by the industrial implementation of optimal control strategy include the unavoidable model plant mismatches and the presence of unknown disturbances. Recently, iterative learning control (ILC) has been used in the run-to-run control of batch processes to directly update input trajectory. The basic idea of ILC is to update the control trajectory for a new batch run using the information from previous batch runs so that the output trajectory converges asymptotically to the desired reference trajectory. Refinement of control signals based on ILC can significantly enhance the performance of tracking control systems. This paper presents a batch to batch iterative learning control strategy for a batch fermentation process using linearised models identified from process operational data. The control policy updating is calculated using multiple linear regression (MLR), partial least squares (PLS) regression or principal component regression (PCR)) model linearised around a reference batch. In order to cope with process variations and disturbances, the reference batch was taken as the immediate previous batch. In such a way, the model is a batch wise linearised model and is updated after each batch. This method has proven to exhibit improving but unsteady results. Therefore, model prediction confidence bounds were incorporated to the control method to improve the results. Steady but slow increments were noticed for a few model prediction confidence bounds penalty factors.

Big Data Analytics for the Redevelopment of Kuala Kedah Jetty

Ganesha Muthkumaran, Nazarudin Mashudi and Ng Kwang Ming
Corporate Technology Division, MIMOS

Corresponding Email: ganesha.muthukumaran@mimos.my

Keywords: *ARIMA, LOESS, Kuala Kedah, Kuala Perlis*

INTRODUCTION

The jetties at Kuala Kedah and Kuala Perlis were to be renovated to upgrade their facilities. At the time of this study, the Kuala Perlis jetty had been upgraded. In line with the upgrade of the Kuala Kedah jetty, plans were made for both jetties to improve the capacity to handle increased passenger loads. This paper will detail the analysis on current people loadings during peak periods via the passenger ticketing information, CCTV and other spatial data relating to the layouts of the jetties. The analysis on the movement of and congregation of people covered includes seating and standing areas as well as retail outlets. This is to ensure a total coverage for proper people traffic management through either jetty.

MATERIALS AND METHODS

Data and Source(s)

Passenger ticketing data for the period 1st Jan 2004 to 1st Dec 2015 for both Kuala Kedah and Kuala Perlis jetties from Jabatan Laut Wilayah Utara was the primary data source used for analysis. CCTV video data of the jetties and the car parks was used for comparison between staff and passengers. Spatial data relating to the layout of the carparks, jetties, seating and standing areas as well as retail outlets was also used.

Analysis

The passenger data was plotted against time for a comparison of the traffic between the two jetties. CCTV video data of passengers from both jetties was used as a means of checking jetty terminal utilization against the passenger ticketing data, and determining the level of staff movement and congregation to establish a baseline noise level. The passenger data was further segregated according to movement from Kuala Kedah to Kuah, Kuah to Kuala Kedah, Kuala Perlis to Kuah and Kuah to Kuala Perlis and plotted against time. The peak periods for passenger traffic at both jetties were determined from these graphs. Peak traffic movements between both jetties were compared and the reasons for these peaks (leisure vs holidays vs work) were elicited through examination of dates (eg: public holidays / weekends vs working days).

Forecasting

An *Autoregressive Integrated Moving Average* (ARIMA) forecast of passenger data from 2016 until 2019 was then produced to determine the trend in passenger traffic through both jetties in Kuala Kedah and Kuala Perlis.

Locally Weighted Scatterplot Smoothing (LOESS) was used for seasonal and trend decomposition.

A daily heatmap of projected utilization levels based on projections of passenger traffic, total departure area (m^2), percentage sitting to be provided, ratio of visitors to passengers, number of installed seats, Level of Service (LoS) measured by area of seating with/without baggage and area of standing with/without baggage was produced.

RESULTS AND DISCUSSION

Graph of total (in and out) passenger data for Kuala Kedah and Kuala Perlis between 1/1/2004 to 1/12/2015 given in Figure 1 below.

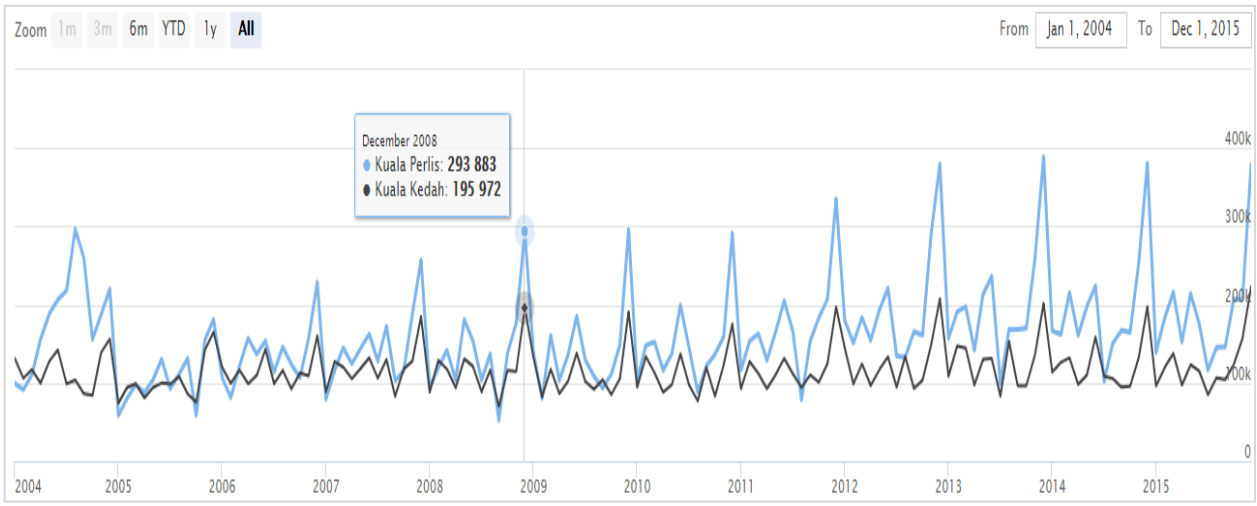


Figure 1: Graph of passenger data for Kuala Kedah and Kuala Perlis.

Both the monthly inbound and outgoing passenger data for both Kuala Kedah and Kuala Perlis were totaled up and graphed against the month the data was collected for. The passenger traffic flow through Kuala Perlis was greater than Kuala Kedah at all times of the year indicating a greater preference to travel through Kuala Perlis after 2008, except for certain very short periods which were determined to be due to the movement of passengers for the purposes of work during the month of Ramadan. The large spikes in the graph appear at the year-end, corresponding with the school holiday period, so the large movement of passengers may be due to travel for leisure.

Figure 2 shows the heatmap of the utilization of the jetty seats by passengers by hour at Kuala Kedah from 27/4/2016 to 5/11/2016

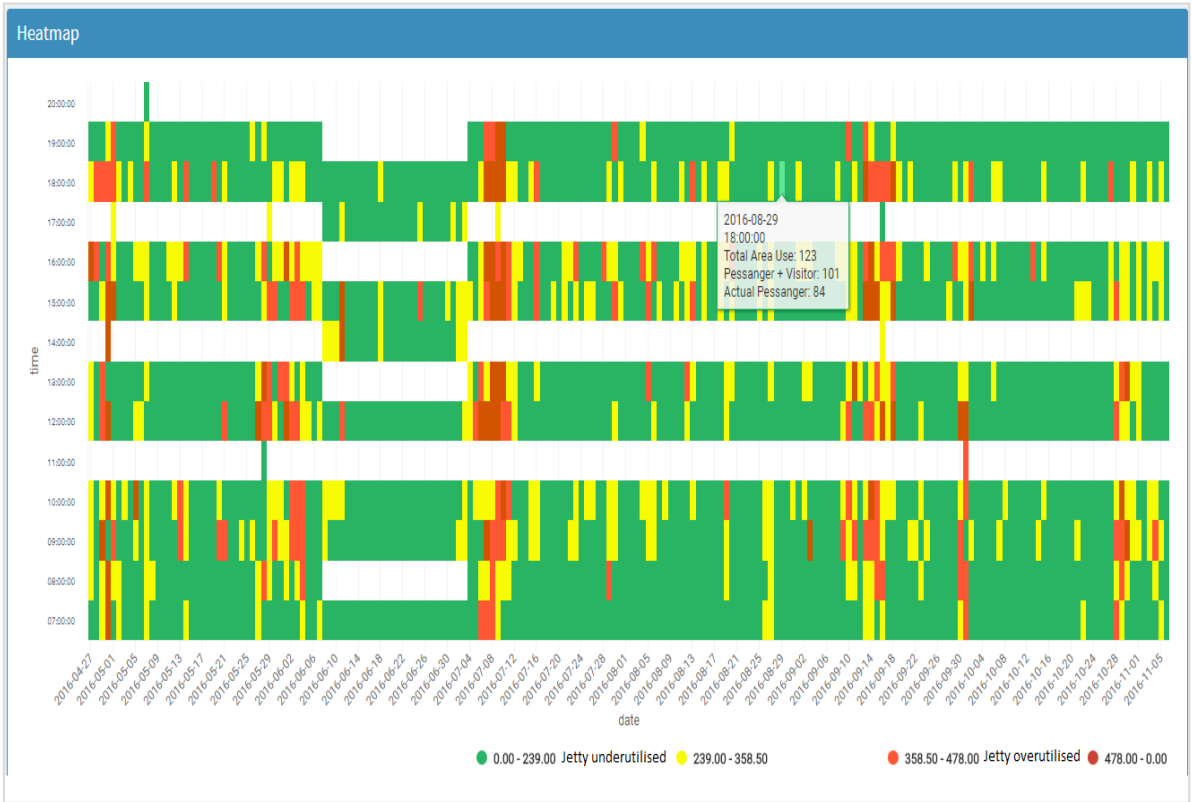


Figure 2: Heat-map of jetty utilization by passengers at Kuala Kedah.

The heat map above was created through analytics using CCTV data in Kuala Kedah jetty. It indicates the utilization of the jetty each hour in terms of the number of passengers present by colour, with green for 0 to 239, yellow for 285 to 427.5, orange for 427.50 to 570.00 and red for numbers exceeding 570.00. Green and yellow mean that the jetty facilities are under-utilized while orange and red mean they are over-utilized.

The ferry terminal at Kuala Kedah is under-utilized for most of the year and passenger traffic only exceeds the existing terminal capacity on peak seasons such as during public holidays, festivals and school holidays, for only short periods.

The heatmap was created only for Kuala Kedah. This was because the jetty in Kuala Perlis had already undergone an upgrade prior to this study on Kuala Kedah.

CONCLUSION

Big Data Analytics was used to study passenger traffic at Kuala Kedah and Kuala Perlis jetties with a view towards improving the former.

Based on the analysis, the following were recommended to the authority concerned:

a) Upgrade of the facilities at Kuala Kedah jetty was necessary.
b) Disseminate appropriate information to passengers to enable better crowd management at Kuala Kedah/Kuala Perlis jetty during peak seasons.
c) Carry out further analysis to identify initiatives to increase passenger traffic flow both ways at Kuala Kedah jetty so as reduce the demand on the Kuala Perlis jetty.

ACKNOWLEDGEMENT

Mr Ng Kwang Ming, Head of Fintech and Systems and Engineering, MIMOS Berhad

En Nazarudin Mashudi, head of project, MIMOS Berhad

REFERENCES

Ministry of Transport, Malaysia big data analytics app developed by MIMOS http://10.4.104.175:3838/mot/app18/