Slangs and Short forms of Malay Twitter Sentiment Analysis using Supervised Machine Learning

Cheng Jet Yin[††], Zakiah Ayop[†], Syarulnaziah Anawar[†], Nur Fadzilah Othman[†], and Norulzahrah Mohd Zainudin [†††]

chengjy580@gmail.com zakiah@utem.edu.my syarulnaziah@utem.edu.my fadzilah.othman@utem.edu.my
norulzahrah@upnm.edu.my

[†]Information Security Forensics and Computer Networking (INSFORNET),
[††]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka (UTeM), 76100, Melaka, Malaysia
[†††]Jabatan Sains Komputer, Fakulti Sains dan Teknologi Pertahanan, Universiti Pertahanan Nasional Malaysia (UPNM), 57000 Kuala Lumpur, Malaysia

Summary

The current society relies upon social media on an everyday basis, which contributes to finding which of the following supervised machine learning algorithms used in sentiment analysis have higher accuracy in detecting Malay internet slang and short forms which can be offensive to a person. This paper is to determine which of the algorithms chosen in supervised machine learning with higher accuracy in detecting internet slang and short forms. To analyze the results of the supervised machine learning classifiers, we have chosen two types of datasets, one is political topic-based, and another same set but is mixed with 50 tweets per targeted keyword. The datasets are then manually labelled positive and negative, before separating the 275 tweets into training and testing sets. Naïve Bayes and Random Forest classifiers are then analyzed and evaluated from their performances. Our experiment results show that Random Forest is a better classifier compared to Naïve Bayes.

Keywords:
Naïve Bayes; Random Forest; Supervised Machine Learning; Twitter

1. Introduction

There are lots of tweets that have been 'ratioed' or 'cancelled' by netizens as they disagree with a certain tweet that is tweeted by a person. The word 'ratioed' is common on Twitter as it meant the negative response or impact that a tweet gets from the public [1]. Cancel culture, also known as call-out culture, is a modern kind of exclusion in which someone is excluded from professional or social circles, either online or in person. The main concept of cancel culture is a variation on the term call-out culture and consists of a form of boycotting or shunning which involves an individual (often a celebrity) who is considered to have shared a questionable opinion or has a problematic behavior that needs the attention of others on social media [2].

Sentiment analysis (opinion mining) is a Natural Language Processing (NLP) technique for evaluating whether data input is neutral, negative, or positive. Sentiment analysis is most performed on textual data to help firms monitor and understand their target market's expectations by analyzing consumer feedback on their brand and products. In this paper, sentiment analysis is used to analyze and rate the tweets found using certain bully trigger words. As everyone has the right to freedom of expression on the Internet, some people have taken advantage of it by using social networking sites like Facebook, Twitter, and Instagram spreading hate speech and criticism, which has led to cybercrime. Some comments or reviews can be beneficial, but some may be insulting or inappropriate to others, as some cybercriminals utilize anonymity or a username to mask their true identity on the Internet. As a result, every form of online bullying could be identified by collecting data from microblogs and social media sites and utilizing sentiment analysis to the data [3].

Most sentiment analysis research used the English language as their main data dictionary, therefore there is a need for a Malay language sentiment analysis study. Furthermore, there is a major challenge in detecting the short forms and internet slang that is widely used in tweets [4]. Internet slang is described as terms or phrases that are often used in online conversations [5]. Short forms, also known as abbreviations, are shortened forms of words or phrases [6].

The remaining parts of the paper are arranged as followed. Section 2 discusses the related research of the sentiment analysis. Section 3 is about the details of the chosen supervised machine learning algorithms. Section 4 is a discussion regarding the results of the chosen classifiers. Lastly, Section 5 will bring closure to our paper and address the scope of future research.

Manuscript received November 5, 2021
Manuscript revised November 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.11.40

2. Related Works

Sentiment Analysis topic in Machine Learning has been studied with many types of methods and algorithms. The three types of approaches will be explained and defined based on their theories and models, which are supervised, unsupervised, and hybrid machine learning.

Under supervised machine learning in sentiment analysis, Sultan [7] proposed an automated sentiment analysis of Afaan Oromoo language utilized by Ethiopian social media users using supervised machine learning. The sentiment can be precisely marked or categorized relying on the true feeling, rather than counting or quantifying polarity terms in sentences or phrases. The model will be effective if it is built on the pattern train and considers the full text rather than just the weight of individual words, according to discourse patterns theory. Reddy et al. [8] created an application to evaluate the sentiment of Twitter data (tweets) expressed by users. The Bayes Theorem employs a probabilistic learning function in this investigation; therefore, the Naïve Bayes (NB) Classifier was utilized. The analyzed tweets in this study have been sorted within the 7 major emotions, ranging from positive, neutral, and negative after using the classifier.

Zammarchi et al. [9] utilized supervised machine learning to explore the temporal evolution of sentiment toward Italy before and during the COVID-19 outbreak using sentiment analysis on Twitter, which analyses changes in the sentiment of tweets in targeted topics and assess the performance of different machine learning classification model to verify the polarity of tweets posted within the period. To explore the polarity of tweets and compare the performance of the Naïve Bayes and Support Vector Machine classifiers that are commonly employed in the analysis of Twitter data gathered in the specified subjects. The Naïve Bayes theory is applied by Lakshmi et al. [10] as a classifier for candidate words generated by the Partially Supervised Alignment Model (PSWAM). The FISTA Algorithm is used to cooperatively construct exact space explicit slant classifiers in most spaces. PSWAM is most commonly employed in sentences to predict the relationships between words to mine opinion relationships.

Meanwhile, for unsupervised machine learning in sentiment analysis, Li et al. [11] developed a novel method for classifying tweet sentiment that includes sentiment-specific word embedding and weighted text characteristics. This was done to demonstrate the utility and efficacy of a newly suggested tweet sentiment classification system that included Tweet Sentiment Scoring Theory, Sentiment-Specific Word Embeddings (SSWE), and a Weighted Text Feature Model (WTFM). This provides a new technique for sentiment analysis on Twitter data, as the orientation must be established firsthand, as well as the appropriate seed/word list as analysis inputs. This algorithm outperforms the two most advanced approaches in tweet sentiment classification.

Batra et al. [12] have compiled Urdu-language tweets and labelled each one with emoticons contained in the tweet text to conduct sentiment analysis. This study necessitates pre-processing, which includes removing columns containing the user's information, the number of retweeted tweets, the account's followers' information, redundant tweets, superfluous punctuation, links, symbols, and spaces, as well as emoji retrieval if present in the tweet text. The final dataset of each tweet record includes columns for 'tweet ID,' text, and emoji derived from the text with an emotion score, as the tweets were translated and the emoticons were analyzed and categorized.

Praveen et al. [13] used sentiment analysis with unsupervised machine learning to analyze the major issues in India as the public discusses issues online, such as the stress, anxiety, and trauma caused by COVID-19. Tweets have been sorted from targeted keywords by positive, neutral, and negative tones. The classifier model used the NLP theory with Textblob, a text library that supports complex analysis and operations on textual data.

Almatarneh and Gamallo [14] suggested an automatic technique for constructing polarity lexicons from corpora, with an emphasis on the generation of a domain-specific lexicon from a corpus of film reviews and the task's use in sentiment analysis. Using the SPLM approach, the Lexicon is built to rank words from negative to positive values. The average of the RF values for two ranges of categories: positive and negative, is then computed. The lexicons were then applied to two datasets of scaled reviews and compared to SO-CALL and SentiWords using the sentiment classifier that was built. The results demonstrate that the SPLM generated automatically outperforms the other lexicons. The list, however, is limited and bound by the domain on which it is described, although it gives scores for a few multi-words.

Zabha et al. [4] presented a lexicon-based technique for cross-lingual sentiment analysis. The system employs both English and Malay lexicons. The data from Twitter was gathered and labelled as positive or negative. A native Malay speaker manually analyzed the data and compared it to the proposed approach to determine correctness. The results have shown that the classifier was able to distinguish the opinions. However, the classifier could not analyze short form and Internet slang.

Lastly for hybrid machine learning for sentiment analysis, Bhowmik et al. [15] implemented a domain-based categorical weighted Lexicon Data Dictionary (LDD) to define sentiment classification from a Bangla language dataset and designed a novel and effective rule-based algorithm to identify sentence polarity classification by retrieving score from a chunk of Bangla text for hybrid machine learning for sentiment analysis. The classifier was able to achieve 82.21% accuracy in cricket topics. However, a specific domain-weighted dictionary word list of adjectives needs to be manually created.

Sentiment analysis was performed by Messaoudi et al. [16] to sort the sentiment of Tunisian tweets. The study examines the importance of several unsupervised word representations (Word2vec, BERT), as well as the use of CNNs and Bidirectional Long Short-Term Memory in Tunisian dialect sentiment analysis on social media. To build the new approach models, the authors employed sentiment lexicon theory but applied CNN with varying filter lengths and a Bidirectional Long Short-Term Memory (Bi-LSTM), a variety of RNNs. This newly built sentiment analysis classifier checks the accuracy and F-score of the texts identified using three initial representations: Word2ves, frWaC, and multilingual BERT (M-BERT) classifiers.

El-Rahman [17] proposed a combination of unsupervised and supervised algorithms to classify specific English tweets. A lexicon-based model is utilized to analyze tweets that have been pre-processed to positive, negative, or neutral. A group of supervised learning models such as Naïve Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME), Decision Tree (DT), Random Forest (RF), and Bagging is applied to train and evaluate the model. Based on the results, Maximum Entropy has the highest accuracy compared to other classifiers.

3. Methodology

3.1 Data Collection

The first type of dataset used in this study is retrieved from Kaggle [18]. It consists of positive and negative Malay language tweets that are on a political topic. This dataset is downloaded and saved into a CSV file. The second type of dataset used consists of targeted short forms and internet slang which can be listed down as a cyberbullying term. These tweets are extracted through Tweepy, by getting developer access to a Twitter account. Keys and tokens are needed to retrieve the data from Twitter API as they act like login credentials. The tweets will be collected, downloaded, and saved into a few CSV files. The targeted keywords that were picked are ‘bodo’ [19][5] and ‘kimak’ [20] as it is considered offensive and bullying.

3.2 Pre-Processing

Tweets will be filtered, cleaned, and pre-processed. Tweets with URLs attached will require a URL remover process. Twitter handles, punctuations, special characters, hashtags, and emojis will be cleaned from the tweets. Stop words will be filtered too. In this case, Indonesian language stop word library is used as it is more complete with the Malay language included due to the similarity of the two languages. After the process is completed, the CSV file is then saved and kept for feature extraction process. All the datasets will be manually labelled into their respective classes, such as positive or negative labels (0 or 1).

3.3 Feature Extraction

Some application features such as TF-IDF will be applied to further clean the tweets by converting tweets into lowercase. The number of times a term appears in a document in relation to the total number of words is calculated using TF, whereas the importance of a phrase is calculated using IDF. This technique can quantify a word in documents, to compute a weight onto each word which highlights the importance of the word. WordCloud is plotted to show the most common, positive, and negative words in the dataset. The datasets are split into two halves, with 80% of the tweets going into the training set and 20% going into the testing set, allowing the machine learning algorithm to learn from the data and generate predictions [21].

3.4 Analysis and Classification

After the above processes are applied to the datasets, Naïve Bayes (NB) and Random Forest (RF) algorithms are applied to determine the performance classes such as the accuracy of the algorithms in detecting short forms and internet slangs that are related to cyberbullying. The comparison of the algorithms will be recorded and analyzed to see which supervised machine learning algorithm is more efficient in carrying out the tasks given. Graphs will be plotted to show a clearer view of the results and another targeted keyword will be picked to determine which supervised machine algorithm is the best.

3.5 Comparison of Classifiers

The comparison of two supervised machine learning algorithms, Naïve Bayes and Random Forest, is performed to see which of the classifiers can detect short forms and internet slang from datasets with the best performance assessors (accuracy and F1-Score). The confusion matrix is utilized to assess the performance of a classification model. It also compares the real values to the machine learning’s model predictions. This will give a detailed perspective of the classifier model’s performance and the variety of errors it makes.

(i) Naïve Bayes

Naïve Bayes is a basic method for building classifier models that can allocate class labels to specific instances, signified as vectors of feature values, with class labels selected from a finite set of labels. Naïve Bayes is a conditional probability model that uses a vector to represent a problem instance for categorization. By utilizing Bayes' theorem, the conditional probability can be represented in Eq. (1) and (2).

$$P(c|x) = \frac{(x|C)P(c)}{P(x)} \quad (1)$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c) \quad (2)$$

(ii) Random Forest

Random Forest algorithm puts together a few algorithms of the same type (multiple decision trees), which gives a result of a forest of trees. The Random Forest algorithm mixes the notions of random subspaces and Bagging. It is a classification and regression ensemble learning technique that produce multiple decision trees throughout the training and provides node output (class) by individual trees. It is sorted into two parts, which are Decision Tree learning and Bagging. It is also a way to average the results of numerous deep decision trees trained with various parts of the same training dataset. The Random Forest algorithm is illustrated in Fig 1.

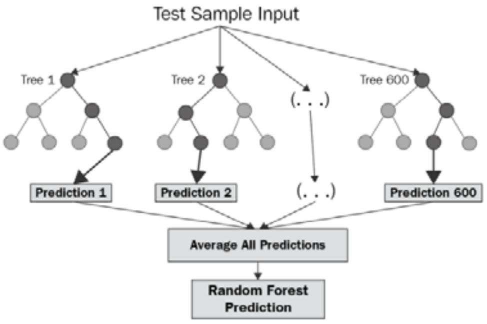


Fig. 1 Random Forest Structure [22].

4. Results and Discussion

4.1 Evaluation Metrics

The confusion matrix, also known as the error matrix, is produced to evaluate the standards of machine learning classifiers. The matrix shows the relationship between the number of correctly or wrongly predicted data and some outputs can be in more than two classes. It has 4 values, which are used to calculate accuracy, precision, recall, and F-measure score (F1). The values are:

(i) True Positive (TP) which shows the number of correctly predicted positive values.

(ii) False Positive (FP) which contains the number of positive values which are wrongly predicted to negative values by the machine learning classifier.

(iii) True Negative (TN) which shows the number of correctly predicted negative values.

(iv) False Negative (FN) which contains the number of negative values which are wrongly predicted to positive values by the machine learning classifier.

Accuracy is calculated as the ratio in the number of correctly predicted data to the total amount of data in the overall dataset. The accuracy formula is shown in Eq. (3).

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (3)$$

Precision is the exactness of the classifier used, known as the percentage of all the data which were predicted to belong in a specific class that was from class. It is also defined as the percentage of the actual positive classes divided by the total of predicted positive classes in the classifier. The precision formula is shown in Eq. (4).

$$precision = \frac{TP}{TP+FP} \quad (4)$$

The recall is defined by the completeness or sensitivity of the machine learning classifier as it is the percentage of data in the dataset that are in the specific class and are labeled correctly to that class. It is also the percentage of the correctly predicted positive values to the actual count of positive data present in the dataset. The recall formula is shown in Eq. (5).

$$recall = \frac{TP}{TP+FN} \quad (5)$$

The F1-score is defined on a per-class basis and is based on the precision and recall results as it is like a balanced average of the two metrics. It also provides a better result of incorrectly classified cases compared to the accuracy metric. The F1-score formula is shown in Eq. (6).

$$F1 = \frac{2 \times precision \times recall}{precision+recall} \quad (6)$$

4.2 Comparative Analysis

Naïve Bayes and Random Forest were used for training and testing the datasets and the outcomes of each classifier were analyzed and compared with each other by referring to the evaluation metrics such as accuracy and F1-Score which are displayed after the training and testing.

Tables 1, 2, and 3 show the analysis results of the supervised machine learning in sentiment analysis using certain short forms and internet slangs found in Malay Twitter as keywords to extract test tweets. 2 types of datasets are used, and 275 tweets are compiled in the test datasets with 55 tweets tested to predict the results. RF has the highest accuracy score, which is 62.22% from the politic topic dataset, 56.36% with the test dataset that contains the keyword ‘bodo’, and 67.27% with the test dataset that contains the keyword ‘kimak’ compared to NB. NB has the least accurate results after the testing, using the datasets given which is 48.89% for the politic topic dataset, 56.36% with keyword ‘bodo’, 56.36% with keyword ‘kimak’. RF has also gotten the highest F1-Score when detecting negative sentiments in short forms and internet slang, which is 65.31% from the politic topic dataset, 63.63% with the test dataset that contains the keyword ‘bodo’, 67.86% with the test dataset that contains the keyword ‘kimak’ in comparison with NB.

Table 1: Comparative Analysis for Politic Topic

Classifiers	Evaluation Parameters	
	Accuracy	F1-Score
NB	48.89 %	51.06 %
RF	62.22 %	65.31 %

Table 2: Comparative Analysis for Targeted Keyword “Bodo”

Classifiers	Evaluation Parameters	
	Accuracy	F1-Score
NB	56.36 %	61.29 %
RF	56.36 %	63.63 %

Table 3: Comparative Analysis for Targeted Keyword “Kimak”

Classifiers	Evaluation Parameters	
	Accuracy	F1-Score
NB	56.36 %	60.00 %
RF	67.27 %	67.86 %

where NB: Naïve Bayes; RF: Random Forest

A new keyword “lancau” is selected to verify the accuracy and F1-Score of RF which makes it better than NB. Table 4 shows the result of the two classifiers using the new Internet slang.

Table 4: Comparative Analysis for Targeted Keyword “Lancau”

Classifiers	Evaluation Parameters	
	Accuracy	F1-Score
NB	61.82 %	70.42 %
RF	69.09 %	77.33 %

From the result, the Random Forest model outperforms the Naïve Bayes model based on the performance indicators, which are accuracy and F1-Score. This is because the Random Forest model is based on a theory that allows it to generate a large number of decision trees and asks each tree to predict the class value. The answer will then be used as the overall prediction, based on the majority vote. During the inputting phase, Random Forest develops multi-altitude decision trees, and the output is in the form of numerous decision trees. By randomly selecting trees, the correlation between trees is reduced which in turn increases the prediction power and efficiency. Predictions are being analyzed by the aggregation of predictions from the various ensemble of datasets used. Random Forest also has lesser variability in the prediction values due to the selection of multiple trees and it handles a large amount of data efficiently. The model requires incremental training to increase the result accuracy and needs recurrent learning with every novel dataset. Previous studies show that the performance of this model is always seen rising, with no downtrend when used with any available datasets [23].

This algorithm also overcomes the overfitting of its classes by providing a bigger dataset, tunning of hyper-parameters, or reducing the number of trees generated in the Random Forest classifier which in turn increases the accuracy of the classifier. Random Forest is versatile as it can be used for both regression and classification tasks. The Naïve Bayes model may be a well-accepted algorithm, but it is outperformed as it is a simple algorithm compared to Random Forest.

The Naïve Bayes model is a fast execution model and uses estimation to get the probability of each word of the document in the product, given the class (likelihood) then multiplied by the probability of the targeted class. Thus, the model will select the class with the highest probability calculated. The accuracy result is variable over time and less time is required to train the classifier, thus lowering the risk of dataset overfitting. Based on the results, the accuracy is low for this model as it has limitations due to the issues with probability calculations, significant variation of human-input text, and the assumptions of independence of predictors. The Random Forest model has a better F1-Score compared to the Naïve Bayes model as it has a better harmonic mean of precision and recall values compared to the other. The Random Forest model has predicted a higher number of correct labels on all labels in the datasets as per the tables above.

In conclusion, the Random Forest classifier shows higher accuracy and F1-Score compared to Naïve Bayes when datasets that contain targeted cyberbullying keywords are used.

5. Conclusion

This research has analyzed the performances of the supervised machine learning algorithms in detecting short forms and internet slang which can be offensive and defined as cyberbullying. Therefore, Random Forest is a better supervised machine learning classifier compared to Naïve Bayes as it has higher accuracy and F1-Score. Another keyword is also picked to extract more tweets to justify that the performance values such as accuracy and F1-scores of Random Forest are higher than Naïve Bayes. New datasets are created in this project to allow new data collection by using the targeted keywords which might be offensive and defined as cyberbullying. Furthermore, this research makes the Internet a safer place as it can detect negative sentiments from offensive tweets even though it contains short forms or internet slang. This can assist cyber security agencies with other social media in the future. The benefits of this study to the Twitter community are the safe environment in the social website to its users and helping other studies with the datasets produced as the Malay language is not widely used in current supervised machine learning for sentiment analysis. Law enforcement agencies can make full use of the results of this research to decide the expressions of the public on social medial platforms. This research contains a few limitations. It is very time-consuming as the Scikit-learn supervised machine learning models take a longer time to train the training dataset, and Spyder will lag when many tweets are being extracted or cleaned. Next, this research is mostly focused on the Malay language, but the extracted tweets might contain other languages such as Indonesian due to the similarities of the two languages. The language library in Python language has only the Indonesian language as more research focused on this language compared to Malay language. This might slightly affect the accuracy of the results. For future works, the usage of the Apache Spark framework can help to increase the size of data extracted without worrying about the lagging problems due to the heavy processes and large time consumption [24]. This research also needs more Malay language library in Python and a compilation of Malay-language tweets and completely labelled datasets to facilitate the pre-processing of the sentiment analysis such as [12].

Acknowledgment

This publication has been supported by Center of Research and Innovation Management (CRIM), Universiti Teknikal Malysia Melaka (UTeM). The authors would like to thank UTeM and INSFORNET research group members for their supports.

References

[1] J. Duribe, “Here’s what being ratioed on Twitter actually means - PopBuzz,” 04-Nov-2020. [Online]. Available: https://www.popbuzz.com/internet/social-media/ratioed-meaning-twitter/. [Accessed: 12-Oct-2021].

[2] L. Mahan, “Youthsplaining: Everything You Need to Know About Cancel Culture - InsideHook,” 20-Aug-2019. [Online]. Available: https://www.insidehook.com/article/internet/youthsplaining-everything-you-need-to-know-about-cancel-culture. [Accessed: 12-Oct-2021].

[3] H. Rosa *et al.*, “Automatic cyberbullying detection: A systematic review,” *Comput. Human Behav.*, vol. 93, pp. 333–345, 2019.

[4] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, “Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, 2019, doi: 10.14569/IJACSA.2019.0100146.

[5] Z. Z. Izazi and T. M. Tengku-Sepora, “Slangs on Social Media: Variations among Malay Language Users on Twitter.,” *Pertanika J. Soc. Sci. ¥& Humanit.*, vol. 28, no. 1, 2020.

[6] “ABBREVIATION | meaning in the Cambridge English Dictionary.” [Online]. Available: https://dictionary.cambridge.org/dictionary/english/abbreviation. [Accessed: 15-Oct-2021].

[7] J. Sultan, “Developing an Automated Machine Learning Based Sentiment Analysis for Afaan Oromoo,” ASTU, 2021.

[8] A. Reddy, D. N. Vasundhara, and P. Subhash, “Sentiment Research on Twitter Data,” *Int. J. Recent Technol. Eng.*, vol. 8, pp. 1068–1070, 2019.

[9] G. Zammarchi, F. Mola, and C. Conversano, “Impact of the COVID-19 outbreak on Italy’s country reputation and stock market performance: a sentiment analysis approach,” *arXiv Prepr. arXiv2103.13871*, 2021.

[10] V. S. Lakshmi, K. Janan, J. P. S. Joshua, and M. Sharoz, “Predicting Supervised Machine Learning Performances for Sentiment Analysis Using Contextual Based Approaches,” in *Journal of Physics: Conference Series*, 2021, vol. 1916, no. 1, p. 12117.

[11] Q. Li, S. Shah, R. Fang, A. Nourbakhsh, and X. Liu, “Tweet sentiment analysis by incorporating sentiment-specific word embedding and weighted text features,” in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 568–571.

[12] R. Batra, Z. Kastrati, A. S. Imran, S. M. Daudpota, and A. Ghafoor, “A Large Scale Tweet Dataset for Urdu Text Sentiment Analysis,” *Mendeley Data*, vol. 1, 2020, doi: 10.17632/RZ3XG97RM5.1.

[13] S. V Praveen, R. Ittamalla, and G. Deepak, “Analyzing

the attitude of Indian citizens towards COVID-19 vaccine--A text analytics study," *Diabetes ¥& Metab. Syndr. Clin. Res. ¥& Rev.*, vol. 15, no. 2, pp. 595–599, 2021.

[14] S. Almatarneh and P. Gamallo, "Automatic construction of domain-specific sentiment lexicons for polarity classification," in *International Conference on Practical Applications of Agents and Multi-Agent Systems*, 2017, pp. 175–182.

[15] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, "Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary," *Nat. Lang. Process. Res.*, vol. 1, no. 3–4, pp. 34–45, 2021.

[16] A. Messaoudi, H. Haddad, M. Ben HajHmida, C. Fourati, and A. Ben Hamida, "Learning word representations for tunisian sentiment analysis," in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2020, pp. 329–340.

[17] S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri, "Sentiment analysis of twitter data," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–4.

[18] H. Zolkepli, "Twitter Political Sentiment in Bahasa | Kaggle," vol. 1. 11-Apr-2018.

[19] "Understanding These Weird Malay Code on Message World - EverydayOnSales.com News." [Online]. Available: https://www.everydayonsales.com/news/understanding-these-weird-malay-code-on-message-world. [Accessed: 19-Oct-2021].

[20] "Malay Slang Wiki | Fandom." [Online]. Available: https://malayslang.fandom.com/wiki/Malay_Slang_Wiki. [Accessed: 19-Oct-2021].

[21] W. Ali, "Phishing website detection based on supervised machine learning with wrapper features selection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 72–78, 2017.

[22] A. Chakure, "Random Forest Regression. In this blog we'll try to understand… | by Afroz Chakure | The Startup | Medium," 29-Jun-2019. [Online]. Available: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f. [Accessed: 19-Oct-2021].

[23] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, and A. Gupte, "Comparative study of classification algorithms used in sentiment analysis," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 5, pp. 6261–6264, 2014.

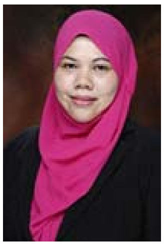
[24] H. Elzayady, K. M. Badran, and G. I. Salama, "Sentiment Analysis on Twitter Data using Apache Spark Framework," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 2018, pp. 171–176.

Cheng Jet Yin received Diploma of IT from Universiti Teknikal Malaysia Melaka in 2020 and currently is a degree student from Universiti Teknikal Malaysia Melaka. Her research interest includes malicious code pattern analysis, data privacy and digital forensics.

Zakiah Ayop holds BSc. in Computer Science (2000) from UTM and MSc in Computer Science (2006) at UPM. Currently she is a senior lecturer in Faculty of Information and Communication Technology (FTMK), Universiti Teknikal Malaysia Melaka (UTeM). She is a member of the Information Security, Digital Forensic, and Computer Networking research group. Her research interest are Information System, Internet of Things (IoT) and Network and Security.

Syarulnaziah Anawar holds her Bachelor of Information Technology (UUM), Msc in Computer Science (UPM), and PhD in Computer Science (UiTM). She is currently a Senior Lecturer at the Faculty of Information and Communication Technology, UTeM. She is a member of the Information Security, Digital Forensic, and Computer Networking (INSFORNET) research group. Her research interests include human-centered computing, participatory sensing, mobile health, usable security and privacy, and societal impact of IoT.

Nur Fadzilah Othman received a degree in Computer Engineering in 2008 and master's in educational technology in 2011 at Universiti Teknologi Malaysia (UTM). In 2017, she obtained her PhD in the field of Information Security at Universiti Teknikal Malaysia Melaka (UTeM). She started her career as a senior lecturer at the Faculty of Information Technology and Communication, UTeM from March 2018. She is an active researcher and has been written and presented a number of papers in conferences and journals.



Norulzahrah Mohd Zainudin is a Lecturer in the Department of Computer Science at Faculty of Defence Science and Technology, National Defence University Malaysia (UPNM). She received her MSc at Universiti Putra Malaysia, BSc at Universiti Teknologi Malaysia and joined Military Academy of Malaysia in 2002. Her main research interests are in the areas of Forensic Computing, Artificial Intelligence Security, IoT Security, and Computer Intelligence. She has published several papers in international journals and conferences. She is a member of Informatics Intelligence Special Interest Group, UPNM and a certified Computer Hacking Forensic Investigator (CHFI) since 2013.