Available online at: https://proceeding.researchsynergypress.com/index.php/cset
RSF Conference Series: Engineering and Technology
ISSN 2809-6843 (Online) | 2809-6878 (Print)
Volume 1 Number 2 (2021): 50-60

Comparative Study of K-Nearest Neighbour and Naïve Bayes Performances on Malay Text Classification

Nazratul Naziah Mohd Muhait[1], Rosmayati Mohemad[2], Noor Maizura Mohamad Noor[3], Zulaiha Ali Othman[4]

[1,2,3]Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu, Malaysia.
[4]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia

Abstract

Police narrative reports are critical in assisting the investigation officer in uncovering hidden information during the criminal investigation process. In recent years, detecting criminal linkages by locating modus operandi in a massive volume of unstructured police reports has become a significant challenge. Here have been few studies on text classification in the Malay language due to some limitations that need to be addressed. Text classification is the process of properly categorizing text into a set of categories. In this study, classification techniques are used to predict the class of modus operandi for housebreaking crime documents using a Malay crime dataset. The dataset used in this study for housebreaking crime is a real dataset from the Royal Police Department of Malaysia. The purpose of this paper is to compare the accuracy of the K-Nearest Neighbour (KNN) and Naive Bayes algorithms for classifying Malay Crime Reports based on their mode of operation. The experiment results show that Naïve Bayes achieved a high accuracy rate of 97.86% with a 9 second execution time, whereas KNN achieved an accuracy rate of 88.43% with a 48 second execution time.

Keywords: Classification, Crime, K-Nearest Neighbour, Naïve Bayes, Malay document

INTRODUCTION

With the extensive usage of the internet nowadays, data sharing has risen proportionately. Due to the growing number of digital documents and the need to analyze, comprehend, organize, and sort them in order to maximize their potential, interest in text document classification has increased in recent years. Text classification (also known as text categorization) is a critical component of text analysis, which is the process of assigning appropriate predefined labels or tags to unstructured text such as phrases, paragraphs, questions, or documents in order to solve a variety of natural language processing problems. It is widely used in a wide range of applications, including spam email detection (Sharma et al., 2021; Ma, Yamamori & Thida, 2020; Taylor & Ezekiel, 2020), sentiment analysis (Waheeb et al., 2020; Alshamsi et al., 2020; Sudhir & Suresh, 2021), question answering (Perevalov & Both, 2021), news categorization (Fanny, Muliono & Tanzil, 2018; Mallick, Mishra & Chae, 2020), and user intent classification (Liu et al., 2020). Classification of pertinent text has demonstrated significant promise in a variety of domains, including marketing, product management, customer service, medical, and others. Text classification also has a significant impact in the criminal domain, as it allows for day-to-day analysis of the crime rate.

Corresponding author
Nazratul Naziah Mohd Muhait, p3938@pps.umt.edu.my
DOI: 10.31098/cset.v1i2.474

Police narrative reports are a valuable source of information for assisting the investigation officer in analyzing and extracting hidden information during the crime analysis process. These reports are a subset of criminal notes written by officers that have narrative construction reflecting the detailed description of the crime's event or incident, including what occurred and why it occurred, in order to focus the investigation's strategy (Quijano-Sánchez et al., 2018). These documents, on the other hand, are notorious for being noisy, resulting in incomplete sentences, misspellings, jargon, mixed use of language, non-regular abbreviations, and other grammar errors (Solomon et al., 2020). Despite the growing number of crimes, the automation of crime reporting system has resulted in massive amounts of unstructured textual crime documents that need to be stored digitally. Detecting criminal linkages within an enormous volume of unstructured police reports has become a significant challenge in recent years (Zhu & Xie, 2019). This detection is advantageous for establishing a pattern of criminal activity committed by the same perpetrator or a criminal group. The outcome may assist the investigation officer in narrowing the area of search. Traditionally, crime linkage detection is accomplished by manually identifying a similar modus operandi pattern based on report analysis. However, owing to the high volume of crimes, this manual procedure is difficult and time intensive, and takes a substantial amount of police resources.

Text classification is one method for addressing these limitations, which can be accomplished through steps such as pre-processing, dimensionality reduction, and categorizing reports based on relevant modus operandi patterns. Based on previous works, there are studies conducted using K-Nearest Neighbour and Naïve Bayes, commonly used classification algorithms, to analyze the pattern of criminal activities (Yadav et al., 2019; Shafi et al., 2021), theft cases categorization (Qi, 2020), and crime prediction (Kim et al., 2019; Almanie, Mirza & Lor, 2015). Most of the studies in text classification have been done for the English text and other well-studied languages. Due to a lack of resources for managing Malay text classification, very few and limited efforts have been carried out for Malay, which varies morphologically and syntactically from other languages. In the criminal domain, on the other hand, text classification has been applied to Malay text to evaluate criminality behaviour (Malim, Sagadevan & Ridzuwan, 2019), crime sentiment analysis (Haron, Abidin & Zamani, 2018), and criminological terms (Lee et al., 2019). Most of these empirical studies focused on Malay social media messages and online news. To the best of the author's knowledge, no research has utilized Malay crime records to infer criminal linkage based on modus operandi categorization.

Therefore, this study explores and compares the performance of K-Nearest Neighbour and Naïve Bayes for classifying Malay housebreaking crime reports. RapidMiner tool was utilized in this work to validate the classification performance of the 1000 Malay housebreaking crime documents in terms of accuracy and timely execution. The best experimental result was obtained using the Naïve Bayes algorithm with the 4-grams of text representation, with 97.86% accuracy rate.

The rest of the paper is organized as follows. The second section discusses related works in text classification, including previous studies on the K-Nearest Neighbour and Naïve Bayes for categorizing unstructured text. Meanwhile, the third section describes the proposed approach for this study. The fourth section summarizes the experimental results and discussion, while the last section outlines the conclusion and this study's future work.

LITERATURE REVIEW

Table 1 illustrates several previous studies in text classification across various domains. The effectiveness of the Naive Bayesian classifier as a robust probabilistic has been demonstrated in the context of completing classification problems efficiently. A few studies using the Naïve Bayes algorithm show a positive outcome, but sometimes it does not work well with the studies. One of

them is the study by Indrayuni (2019), which used Naive Bayes combined with a 2-gram model and achieved a high accuracy of 90.50% with an area under the curve (AUC) value of 0.715. The study aims to categorize customer feedback as negative or positive. Sánchez-Franco, Navarro-Garcia, and Rondán-Catalua (2019) conducted a similar study in which they used Naïve Bayes to automatically classify customer satisfaction from massive volumes of English-based customer reviews into specific positive and negative reviews. The experiment yielded significant results, with most reviews were correctly classified with an AUC value of 0.86 and a high precision and recall score of greater than 0.84. Both studies demonstrate that when combined with the n-gram feature, the Naive Bayes algorithm can accurately categorize user reviews as positive or negative for sentiment analysis. Meanwhile, a comparative study of Naïve Bayes and Support Vector Machine (SVM) for classifying student complaints reveals contrast findings. SVM achieves higher accuracy and AUC value of 84.45% and 0.922, respectively than Naïve Bayes, which obtains only 69.75% accuracy and 0.679 for AUC values (Hermanto, Mustopa & Kuntoro, 2020). Cross-validation, confusion matrix, and receiver operating characteristic (ROC) curve are used to compare the findings. Based on these findings, it can be concluded that applying the Naïve Bayes method does not always result in improved accuracy and that it also relies on the data kinds of documents. Watmah, Suryanto & Martias (2021) conducted a Shopee review classifier using K-Nearest Neighbour, support vector machine, and random forest. The objective of this study is to classify the customer review at the play store application. The comparison results show that the support vector machine (SVM) is the best classifier with 89.4% accuracy, 89.5% precision, and 89.7% recall. There is only a slight difference in the accuracy between KNN and SVM classifier by 0.40 %. KNN accuracy, precision, and recall is 89.0%, 89.7%, and 87.5%. This means KNN also has its own strengths and may need to use appropriate parameters to overcome SVM performance. Random forest is also quite good, with an accuracy value of more than 80% in this study. Random forest classifier achieves 83.0% for accuracy, 85.7% for precision, and 81.4% for recall. The other work with the same language of Indonesia is done by Khoirunnisa et al. (2020) to investigate the effect of using the N-gram model in document classification. She used only one type of classifier, which are Naïve Bayes, but she added up with the N-gram model in text processing. The purpose why this study using N-gram is to booster the performance of classification. However, the implementation of the N-gram was not bringing the happiness to them because the performance become lower than applying N-gram. Finally, they have decided to do the classification task without using N-gram and they obtain the impressive performance with 84.97% of accuracy rather than 32.68% of accuracy with N-gram. Other than the business and education field, the crime domain also uses classification techniques to define the crime category for different states (Iqbal *et al.*, 2013). There are two types of algorithms that perform classification tasks in this study which are Naïve Bayes and Decision Tree. The experiment uses the real dataset from 1995 FBI UCR to predict the crime category for different states of USA. As a result, the Decision Tree performs well than Naïve Bayes for accuracy, precision, and recall value with 83.95%, 83.5%, and 84%. The accuracy, precision and recall for Naïve Bayes is 70.81%, 66.4%, and 70.8%. Most of the studies use datasets containing one language of the document, but there are some studies that use datasets containing two languages of the document to test the effectiveness of the classification task. Jaafar, Indra & Zamin (2016) have developed one study that classifies the textual documents that have the same morphology, which is Indonesia and Malaysia. The experiment was conducted to classify the 280 news of Malaysia and 280 news of Indonesia documents between the years 2014 and 2015. K-Nearest Neighbor was implemented in this study to see the effectiveness of this algorithm in classifying languages and document categories. The classifier produces the best result for accuracy rate with 95.63% and accuracy for language with 97.50%. There are also studies that use Twitter datasets for the classification task. Tiun (2018) are using Twitter data to classify the negative and positive for the Malay short text. The study uses three types of classifiers includes K-nearest neighbour, support vector machine, and Naïve Bayes. The result shows that support vector machines get higher accuracy with 95%. Support

vector machine are suitable to be used if the dataset contains two class or binominal types. It is not supported if the dataset contains more than two class.

Table 1: Summary of previous studies using classification techniques.

Reference	Domain	Technique / Algorithm	Data Type	Language	Objective	Result
(Watmah, Suryanto & Martias, 2021)	Business	K-Nearest Neighbour, Support Vector Machine, Random Forest	Customer review	Indonesia	identify user satisfaction	SVM,89.4 % accuracy, 89.5% precision, 89.7% recall
(Khoirunnisa *et al.*, 2020)	Mass media	N-gram + Naïve Bayes	News document	Indonesia	identify effect of N-gram on document classification	Applying n-gram 32.68% accuracy, not apply 84.97% accuracy.
(Hermanto, Mustopa & Kuntoro, 2020)	Education	Naïve Bayes, Support Vector Machine	Students complain	Indonesia	classify student complain	SVM: Accuracy = 84.45% AUC = 0.922 Naïve Bayes Accuracy = 69.75% AUC = 0.679
(Indrayuni, 2019)	Business	N-gram + Naïve Bayes	Product review	Indonesia	classify negative & positive review	Accuracy = 90.50% AUC = 0.715
(Sánchez-Franco, Navarro-Garc'ia & Rondán-Cataluña, 2019)	Business	N-gram + Naïve Bayes	Customer review	English	classify negative & positive review	Accuracy = 84% AUC = 0.86

(Jaafar, Indra & Zamin, 2016)	Mass media	K-Nearest Neighbour	Online news documents	Malay & Indonesia	identify news language	Language 95.63%, category 97.5%
(Iqbal *et al.*, 2013)	Crime	Naïve Bayes, Decision Tree	Crime dataset	English	define crime category for different state.	NB 83.955% accuracy

RESEARCH METHOD

The figure below shows the research framework of text classification for this study. As shown, the framework consists of five phases: document collection, text pre-processing, document representation, modus operandi classification, and analysis and evaluation of classification.

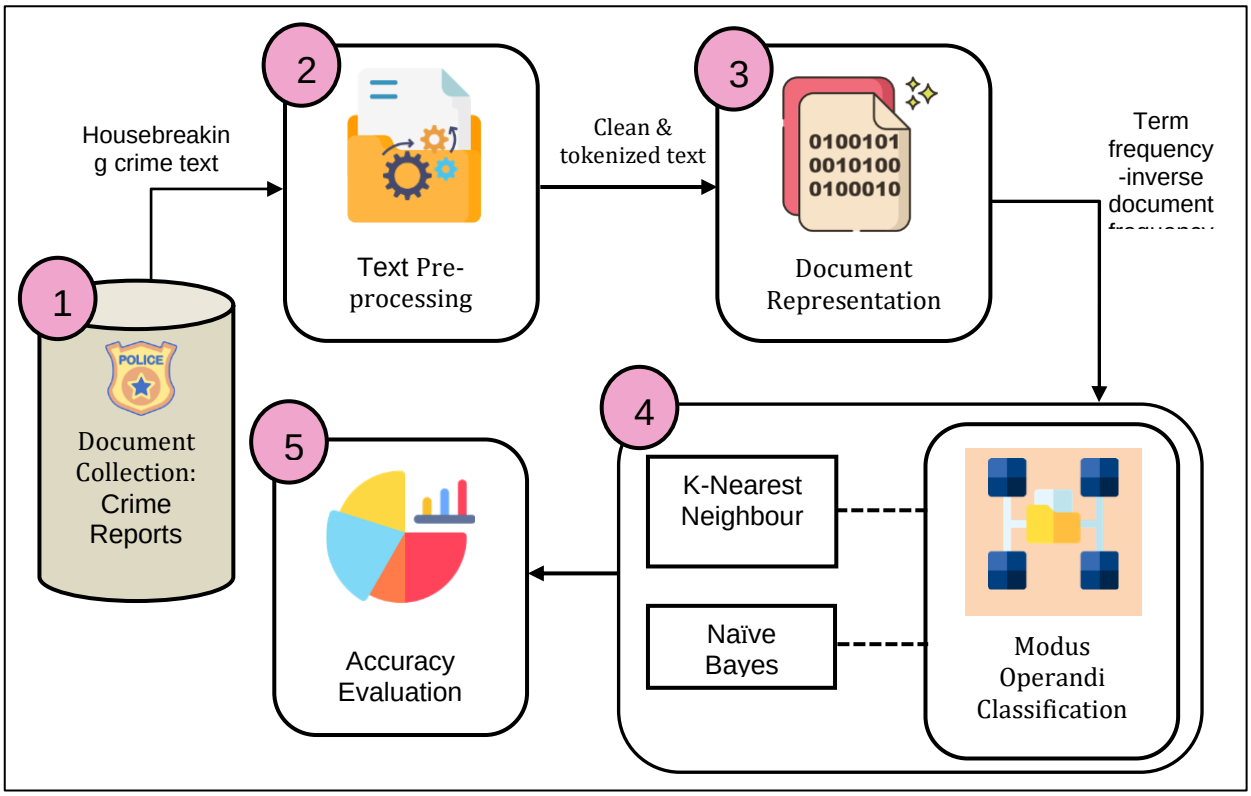

Figure 1: Research method framework

a) Document Collection

The experiment utilized a dataset of housebreaking crime records from 2010 to 2013. High-quality research datasets are carefully selected since past research has shown that low-quality datasets have a negative impact on machine learning findings, especially on machine learning accuracy (Botsis *et al.*, 2010; Dai & Berleant, 2021). It is a closed domain dataset collected from the Royal Police Department of Malaysia. The corpus contains 100,383 Malay crime reports. For the experiment, 1,000 crime reports were randomly chosen and screened using five distinct modus

operandi class: *cara* (method), *peranan* (role), *keganjilan* (oddity), *senjata* (weapon), and *tempat* (location). he crime investigator has manually identified the modus operandi of each crime report. The number of documents for each modus operandi category is evenly distributed. The document sizes ranged from 1 to 145 words, with a total word count of 11,524. Around 72 documents are blank, indicating that they are part of a noise dataset. These documents are then replaced with non-empty crime reports at random.

b) Text Pre-processing

Text pre-processing is used to remove noisy and language-dependent elements from the unstructured text. Pre-processing includes tokenization, case transformation, stop word removal, stemming, token length filtering, and n-gram generation. This stage is critical to removing unnecessary textual elements from documents, and hence increasing classification efficiency, accuracy, and speed.

Tokenization is the process of breaking down a large body of text into a single word, phrase, or symbol. The purpose of tokenization is to examine each sentence in the document and then to identify the document's keyword. **Case transformation**, on the other hand, is the process of converting text to lowercase characters. Since all the text in the housebreaking crime reports is in uppercase, this straightforward step is crucial for reducing the dimensionality of data and significantly improving expected output consistency. **Stop word removal** is a method to eliminate insignificant words from textual documents. These kinds of words are frequently derived from pronouns, prepositions, conjunctions, numbers, punctuation marks, and symbols. Typically, these terms are less meaningful and thus irrelevant for classification purposes. Meanwhile, **stemming** is used to determine the root word. Linguistically, the root word is referred to as a morpheme, which is the smallest unit of words that retain their meaning and cannot be further subdivided. Due to the diversity of morphological structures in Malay, chunking the word to its root is beneficial for reducing the dimensionality of data and allowing classification algorithms to work more effectively. The following step is **filtering token by length**. It is used to specify the minimum, and maximum character counts that should be recognized. **Generating n-grams** is also critical in this phase in order to achieve high classification accuracy. Occasionally, each token does not have a distinct meaning, and thus adding n-gram to the document representation may improve classification performance. In this study, the feature is extracted using a 4-gram (term) because it produces the highest accuracy when compared to 1-gram, 2-gram and 3-gram. The stemming and stop word removal processes are carried out independently, as RapidMiner tools lack an operator for stemming Malay text. Figure 2 shows the illustration for text pre-processing in RapidMiner.

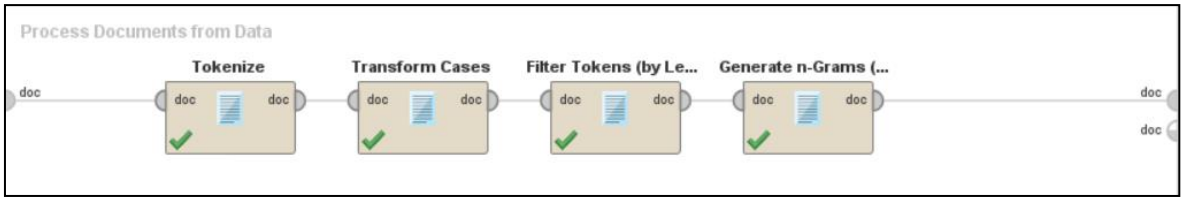

Figure 2: Text pre-processing in RapidMiner

c) Document representation

Classification algorithms are incapable of directly comprehending textual documents. Because of this, the documents need to be transformed into numerical representation before it can be processed by classifiers. The documents in this study are represented using a vector space model. Frequency (TF-IDF) matrix. The term in a document is determined using TF-IDF. The values of the

terms range from 0 to 1, with 0 denoting insignificant terms and 1 denoting significant terms. The example of TF-IDF is shown in Table 2.

Table 2: Term frequency-inverse document frequency values

	abdullah	abu	acu	ada	adik	agama
Document 1	0	0	0	0	0	0.325
Document 2	0	0	0	0	0.683	0
Document 3	0	0	0	0	0.362	0
Document 4	0	0	0	1	0	0
Document 5	0	0	0.176	0	0	0
Document 6	0	0.497	0	0	0	0
Document 7	0.424	0	0	0	0	0
Document 8	0.344	0	0	0	0	0

d) Classification Algorithms

Document classification is a two-step process. The training process is the first step, during which the model is created. The training process is the process by which the algorithm learns to identify characteristics in examples that enable objects of different classes to be distinguished. The classification process is the next step. This step predicts the document's class label using the previously created model. The accuracy of a classifier is defined as the percentage of correctly recognized test data. This study employed two distinct algorithms for classification: K-nearest neighbour and Naïve Bayes.

Figure 3 depicts the overall classification process used by RapidMiner. According to the figure, the first step is to load the dataset into RapidMiner, where the appropriate operator is selected to perform the classification. Depending on our data types, we can choose from a variety of operators to carry out the process. Due to the unstructured nature of the documents in this study, the nominal to text operator is used. Prior to that, the document's role must be defined in order to specify which attributes to predict. Then, the process document from the data operator is used to perform the pre-processing of the text. The data is then partitioned into 700 documents for training and 300 documents for testing in a 70:30 ratio. Following that, the classification algorithm operator is connected to the apply model operator, and finally, the performance operator is selected. There are numerous types of performance operators depending on the task, but for the purposes of this study, the performance (classification) operator is used exclusively for classification.

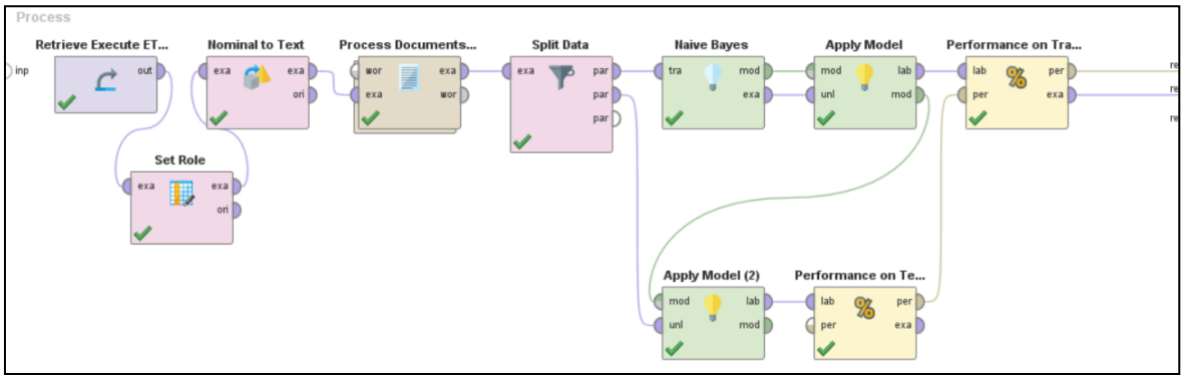

Figure 3: Classification process in RapidMiner

e) Accuracy Evaluation

Finally, the classifier's performance is evaluated in terms of accuracy, precision, recall, and execution time. The precision metric indicates the percentage of correct classifier predictions, whereas the recall metric indicates the percentage of items in a class correctly identified by the classifier.

FINDINGS AND DISCUSSION

The experimental study compares the K-Nearest Neighbour and Naive Bayes classification algorithms. There are 1000 Malay datasets to import into the RapidMiner tool. Both algorithms are tested independently using 10-fold cross-validation and a mixed measure type for a total of five classes. The accuracy of 4-gram is 88.43% for K-Nearest Neighbour and 97.86% for Naïve Bayes. As a result, Naïve Bayes outperforms K-Nearest Neighbour and demonstrates superior performance. Additionally, Table 3 and Table 4 illustrate the confusion matrixes for these two algorithms.

Table 3: Confusion matrix using K-Nearest Neighbour

Accuracy: 88.43%							
class	true *cara*	true *keganjilan*	true *peranan*	true *senjata*	true *tempat*	class precision	total
pred.Cara	131	9	22	1	0	80.37%	163
pred.Keganjilan	1	113	4	0	2	94.17%	120
pred.Peranan	5	11	108	3	2	83.27%	129
pred.Senjata	2	4	4	132	1	92.31%	143
pred.Tempat	1	3	2	4	135	93.10%	145
class recall	93.57%	80.71%	77.14%	94.29%	96.43%		

Table 4: Confusion matrix using Naive Bayes

Accuracy: 97.86%							
class	true *cara*	true *keganjilan*	true *peranan*	true *senjata*	true *tempat*	class precision	total
pred.Cara	139	0	0	0	0	100.00%	139
pred.Keganjilan	0	140	0	0	0	100.00%	140
pred.Peranan	1	0	134	0	0	99.26%	135
pred.Senjata	0	0	6	140	8	90.91%	154
pred.Tempat	0	0	0	0	132	100.00%	132
class recall	99.29%	100.00%	95.71%	100.00%	94.29%		

The classification results in Table 3 were obtained using the K-nearest neighbour algorithm. There are five pre-labelled classification categories: *cara* (method), *keganjilan* (oddity), *peranan* (role), *senjata* (weapon), and *tempat* (location). Around 163 documents are classified as *cara*, with 131 of those documents correctly classified. Meanwhile, the remainder of the documents are incorrectly categorized as *keganjilan*, *peranan*, *senjata*, and *tempat*. In addition, there are 120 documents classified as *keganjilan*. Among these 120 documents, 113 are correctly classified as *keganjilan*, while the remaining seven are incorrectly classified. There are 129 documents classified as *peranan*

and 108 documents that are correctly labelled. Furthermore, from 143 documents that have been classified as *senjata*, around 132 documents are classified correctly as *senjata*. Finally, out of 145 documents, class *tempat* contains 135 that have been correctly classified. Table 4 illustrates the confusion matrix for the Naïve Bayes algorithm. The Naïve Bayes algorithm was found to be the most effective classifier in this study because it classified the majority of documents correctly into their respective classes. The majority of classes achieve 100% precision.

Table 5 compares the k-nearest neighbour and Nave Bayes results. Naive Bayes performed admirably, achieving higher accuracy, precision, recall, and execution time in as little as 9 seconds. With the lengthy execution time, K-nearest neighbour only achieves 88.43% accuracy. It can be concluded that the best classifier for classifying Malay text documents is Nave Bayes.

Table 5: Accuracy, Incorrected Classified documents, Precision, Recall and time execution for both Algorithms

Algorithm	Accuracy (Correctly Classified documents)	Incorrected Classified documents.	Precision	Recall	Time execution (second)
K-Nearest Neighbour	88.43%	11.57%	88.73%	88.43%	48
Naïve Bayes	97.86%	2.14%	98.03%	97.86%	9

CONCLUSION

The purpose of this study is to conduct a comparison of the K-Nearest Neighbour and Naive Bayes algorithms. The experiment is designed to predict modus operandi classes using predefined label documents as input. This research's primary objective is to aid investigators in identifying and comprehending specific and current trends in housebreaking crime. This study analyzed a real-world dataset from the Malaysian Royal Police Department between 2010 and 2013. The primary objective of this research is to develop a predictive model that accurately predicts documents based on their modus operandi. The algorithms K-nearest neighbour and Naive Bayes were used to evaluate the performance accuracy and execution time in this study. Other classification models will be investigated further in the future to improve crime prediction accuracy and overall performance as a continuation of our work.

ACKNOWLEDGMENT

This research is supported by Ministry of Education Malaysia, under Fundamental Research Grant Scheme (FRGS) with vote number 59541 (Reference Code: FRGS/1/2018/ICT04/UMT/02/3). The authors would also like to acknowledge Royal Police Department of Malaysia for their full support of this research

REFERENCES

Almanie, T., Mirza, R. & Lor, E. (2015) Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. International Journal of Data Mining & Knowledge Management Process. [Online] 5 (4), 01–19. Available from: doi:10.5121/ijdkp.2015.5401.

Alshamsi, A., Bayari, R., Salloum, S. & others (2020) Sentiment analysis in English texts. Advances in Science, Technology and Engineering Systems Journal. 5 (6), 1683–1689.

Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. (2010) Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics. 2010, 1.

Dai, W. & Berleant, D. (2021) Benchmarking Deep Learning Classifiers: Beyond Accuracy. arXiv preprint arXiv:2103.03102.

Fanny, F., Muliono, Y. & Tanzil, F. (2018) A comparison of text classification methods k-NN, Na{\"\i}ve Bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT. 3 (2), 157–160.

Haron, M.B.C., Abidin, S.Z.Z. & Zamani, N.A.M. (2018) Visualization of Crime News Sentiment in Facebook. International Journal of Engineering \& Technology. 7 (4.38), 955–959.

Hermanto, H., Mustopa, A. & Kuntoro, A.Y. (2020) Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). [Online] 5 (2), 211–220. Available from: doi:10.33480/jitk.v5i2.1181.

Indrayuni, E. (2019) Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. Jurnal Khatulistiwa Informatika. [Online] 7 (1), 29–36. Available from: doi:10.31294/jki.v7i1.1.

Iqbal, R., Murad, M.A.A., Mustapha, A., Panahy, P.H.S., et al. (2013) An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology. [Online] 6 (3), 4219–4225. Available from: doi:10.17485/ijst/2013/v6i3.6.

Jaafar, J., Indra, Z. & Zamin, N. (2016) A category classification algorithm for Indonesian and Malay news documents. Jurnal Teknologi. [Online] 78 (8–2), 121–132. Available from: doi:10.11113/jt.v78.9549.

Khoirunnisa, F., Yusliani, N., T, M., Rodiah, D., et al. (2020) Effect of N-Gram on Document Classification on the Naïve Bayes Classifier Algorithm. 01 (01), 26–33.

Kim, S., Joshi, P., Kalsi, P.S. & Taheri, P. (2019) Crime Analysis Through Machine Learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018. [Online] (January), 415–420. Available from: doi:10.1109/IEMCON.2018.8614828.

Lee, J.C.L., Teh, P.L., Lau, S.L. & Pak, I. (2019) Compilation of malay criminological terms from online news. Indonesian Journal of Electrical Engineering and Computer Science. [Online] 15 (1), 355–364. Available from: doi:10.11591/ijeecs.v15.i1.pp355-364.

Liu, H., Liu, Y., Wong, L.-P., Lee, L.-K., et al. (2020) A Hybrid Neural Network BERT-Cap Based on Pre-Trained Language Model and Capsule Network for User Intent Classification. Complexity. 2020.

Ma, T.M., Yamamori, K. & Thida, A. (2020) A Comparative Approach to Naive Bayes Classifier and Support Vector Machine for Email Spam Classification. In: 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). 2020 pp. 324–326.

Malim, N.H.A.H., Sagadevan, S. & Ridzuwan, N.I. (2019) Criminality Recognition using Machine Learning on Malay Language Tweets. Pertanika Journal of Science \& Technology. 27 (4).

Mallick, P.K., Mishra, S. & Chae, G.-S. (2020) Digital media news categorization using Bernoulli document model for web content convergence. Personal and Ubiquitous Computing. 1–16.

Perevalov, A. & Both, A. (2021) Improving Answer Type Classification Quality Through Combined Question Answering Datasets. In: International Conference on Knowledge Science, Engineering and Management. 2021 pp. 191–204.

Qi, Z. (2020) The text classification of theft crime based on TF-IDF and XGBoost model. In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). 2020 pp. 1241–1246.

Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J. & Camacho-Collados, M. (2018) Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. Knowledge-Based Systems. 149, 155–168.

Sánchez-Franco, M.J., Navarro-Garc'ia, A. & Rondán-Cataluña, F.J. (2019) A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. Journal of Business Research. 101, 499–506.

Shafi, I., Din, S., Hussain, Z., Ashraf, I., et al. (2021) Adaptable reduced-complexity approach based on state vector machine for identification of criminal activists on social media. IEEE Access. 9, 95456–95468.

Sharma, V.D., Yadav, S.K., Yadav, S.K., Singh, K.N., et al. (2021) An effective approach to protect social media account from spam mail--A machine learning approach. Materials Today: Proceedings.

Solomon, A., Magen, A., Hanouna, S., Kertis, M., et al. (2020) Crime Linkage Based on Textual Hebrew Police Reports Utilizing Behavioral Patterns. In: Proceedings of the 29th ACM International Conference on Information \& Knowledge Management. 2020 pp. 2749–2756.

Sudhir, P. & Suresh, V.D. (2021) Comparative study of various approaches, applications and classifiers for sentiment analysis. Global Transitions Proceedings.

Taylor, O.E. & Ezekiel, P.S. (2020) A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier. Int. J. Comput. Sci. Math. Theory. 6, 1–11.

Tiun, S. (2018) Experiments on Malay short text classification. Proceedings of the 2017 6th International Conference on Electrical Engineering and Informatics: Sustainable Society Through Digital Innovation, ICEEI 2017. [Online] 2017-Novem (i), 1–4. Available from: doi:10.1109/ICEEI.2017.8312371.

Waheeb, S.A., Ahmed Khan, N., Chen, B. & Shang, X. (2020) Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. Information. 11 (5), 281.

Watmah, S., Suryanto & Martias (2021) Komparasi Metode K-NN, Support Vector Machine, dan Random Forest pada E-Commerce Shopee. INSANtek – Jurnal Inovasi dan Sains Teknik Elektro. 2 (1), 15.

Yadav, N., Kumar, A., Bhatnagar, R. & Verma, V.K. (2019) City crime mapping using machine learning techniques. In: International Conference on Advanced Machine Learning Technologies and Applications. 2019 pp. 656–668.

Zhu, S. & Xie, Y. (2019) Spatial-temporal-textual point processes with applications in crime linkage detection. arXiv preprint arXiv:1902.00440.