International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 4, August 2021, pp. 3617~3628
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i4.pp3617-3628

Text classification model for methamphetamine-related tweets in Southeast Asia using dual data preprocessing techniques

Narongsak Chayangkoon, Anongnart Srivihok
Department of Computer Science, Faculty of Science, Kasetsart University, Thailand

Article Info

Article history:

Received Aug 28, 2020
Revised Jan 4, 2021
Accepted Jan 18, 2021

Keywords:

Data preprocessing
Feature selection
Methamphetamine
Text classification
Tweet

ABSTRACT

Methamphetamine addiction is a prominent problem in Southeast Asia. Drug addicts often discuss illegal activities on popular social networking services. These individuals spread messages on social media as a means of both buying and selling drugs online. This paper proposes a model, the "text classification model of methamphetamine tweets in Southeast Asia" (TMTA), to identify whether a tweet from Southeast Asia is related to methamphetamine abuse. The research addresses the weakness of bag of words (BoW) by introducing BoW and Word2Vec feature selection (BWF) techniques. A domain-based feature selection method was performed using the BoW dataset and Word2Vec. The BWF dataset provided a smaller number of features than the BoW and TF–IDF dataset. We experimented with three candidate classifiers: Support vector machine (SVM), decision tree (J48) and naive bayes (NB). We found that the J48 classifier with the BWF dataset provided the best performance for the TMTA in terms of accuracy (0.815), F-measure (0.818), Kappa (0.528), Matthews correlation coefficient (0.529) and high area under the ROC Curve (0.763). Moreover, TMTA provided the lowest runtime (3.480 seconds) using the J48 with the BWF dataset.

Corresponding Author:

Narongsak Chayangkoon
Department of Computer Science, Faculty of Science
Kasetsart University
Bangkok 10900, Thailand
Email: narongsak.chay@ku.th

1. INTRODUCTION

Southeast Asia is considered a centre of methamphetamine production because of many related arrests, which continue to rise annually after increasing four-fold from 1998 to 2014 [1]. Drug addicts often talk about activities related to methamphetamine on popular social networking services. Some tweets are published on social media for the purposes of buying and selling drugs online. However, little research has examined the development of text classification models for tweets relating to methamphetamine [2]. This study's objective is to propose a new data preprocessing technique for methamphetamine-related tweets in Southeast Asia.

For this purpose, we have introduced a model called the "text classification model of methamphetamine tweets in Southeast Asia using dual data preprocessing techniques (TMTA)". A critical process in the development of the TMTA was data preprocessing using the bag-of-words (BoW) model, a basic, classical, straightforward technique, popular for data preprocessing in text classification. This method considers the frequency of each word as a classification feature known as one-hot representation. Each word is represented by a sparse vector consisting of its index and frequency [3, 4]. As features may potentially run

into multiple vectors, Bow's weakness is the likelihood of resulting in a larger size in the form of high-dimensional vectors [5]. The current study reduced this weakness of BoW by proposing the BWF technique, a novel approach consisting of BoW and Word2Vec feature selection comprising two steps. The first step created a text representation dataset using BoW. The second involved a domain-based feature selection, performed using the BoW dataset and Word2Vec.

We began by collecting tweets from Twitter that originated from Southeast Asia and dividing them into two classes, namely abuse and non-abuse. We then experimented with three classification algorithms, including support vector machine (SVM), decision tree (J48) and naive bayes (NB). We measured the performance of each model based on accuracy, F-measure, the Area under the ROC curve (AUC), Kappa, Matthews correlation coefficient (MCC) and runtime. Finally, we compared our model with three different data preprocessing techniques (BoW, TF–IDF, BWF). The experimental results showed that the TMTA, using the J48 and BWF dataset, provided the highest performance measurements.

This research contributes to the literature a new data preprocessing technique for classifying methamphetamine-related tweets. BWF provides a smaller dataset than traditional or widely used techniques such as BoW and TF–IDF. Furthermore, the TMTA model can accurately identify narcotic methamphetamine tweets. Hence, this model can be developed as an application system to monitor tweets related to methamphetamine on the Twitter platform in Southeast Asia.

Although a handful of researchers have used different classifiers to develop text classification models for tweets related to illegal drugs, few research studies are available. Phan *et al.* [2] developed a model to detect the sharing of tweets related to illegal drugs, including marijuana, cocaine and heroin. The authors conducted their research in a rural region of the United States of America (USA). Their dataset was divided by experts into 2 classes: Abuse or non-abuse. BoW and TF-IDF were used as data preprocessing techniques, and 3 classifiers were used: SVM, J48 and NB. The study findings revealed that the best model was the J48 algorithm using the TF–IDF method, which provided the highest F-measure of 0.7480. Ragini and Anand [6], in a study addressing the multi-class classification problem for a disaster event in India, collected 70,817 relevant tweets from 2014 to 2015. They divided the tweets into 7 classes: food, water, shelter, and medical emergency, people trapped, collapsed structure and electricity. Next, the authors created models using SVM and NB classifiers. The best-performing model in this case used the SVM classifier with the TF-IDF dataset. Wang *et al*. [7] compared the efficiency of data preprocessing techniques consisting of BoW, TF–IDF, PV-DM and PV-DBOW. The dataset used in the experiment, based on the Shanghai and Shenzhen Stock Exchanges, was divided into 2 datasets: small class and big class. The classification models were NB, logistic regression, SVM, K-nearest neighbour (KNN) and Decision Tree. The researchers reported that the small class dataset, using the SVM algorithm with the TF-IDF dataset, demonstrated the highest accuracy of 0.8355.

Ghosh *et al*. [8] addressed the multi-class classification problem for disaster events consisting of earthquakes, hurricanes, electrical outages and drought. The experimental tweets in the 2015 dataset related to the Nepal earthquake in April of that year. The TF–IDF method provided the dataset for the models that were created using the following classifiers: NB, SVM, Decision Tree, AdaBoost, random forest and gradient boosting. According to the results, the model created using SVM with the TF-IDF dataset provided the highest F-measure of 0.9178. Burel and Alani [9] also addressed disaster events with a dataset that consisted of 28,000 tweets on various crises between 2012 and 2013. Their two models were based on the convolutional neural network (CNN) classifier using a word-embedding dataset and the SVM classifier using the TF-IDF dataset. The results showed that CNN with a word-embedding dataset did not significantly outperform SVM with the TF-IDF dataset. The literature review also covers classifiers for the development of text classification models using tweeted data with classifiers consisting of SVM, J48 and NB. The SVM classifier with TF–IDF was widely used to develop the text classification model. Additionally, researchers chose the J48 and NB classifiers to develop the text classification model.

Text representation is a part of natural language processing (NLP), which converts text data into numeric vectors that the machine can manipulate. Numerous methods can perform text data conversion. One simple approach gives each word a one-hot representation, such as BoW. In addition, TF-IDF text representation is a popular technique for developing a text classification model. As mentioned, BoW involves a collection of words that represents the features of the text by the word frequency. For example, a word has a value of one if it appears once in the text. The vector representation of text using BoW is an unstructured text document [3, 4]. Furthermore, term frequency (TF) is a calculation of the frequency of a word that appears in the document relative to the total number of words in the document. A high TF value indicates the importance of the word. In addition, inverse document frequency (IDF) is the inverse of the word frequency in the document. A high IDF value indicates an important word, which should appear only in that category and not in other categories. Therefore, term frequency-inverse document frequency (TF-IDF) is the weight

indicating the importance of the word. TF-IDF determines the weight of the word (*w*) in a document (*d*) that appears in the document, based on (1) [10, 11].

$$TF - IDF = (TFw, d \times IDFw) \quad (1)$$

Tomas Mikolov developed Word2Vec as a tool for NLP. This tool, which employs deep neural networks that train word associations to synonymous words, is used to create a pre-trained word embedding model that is trained from the corpus. Word2Vec has two different algorithms: The Skip-gram model and continuous bag-of-words (CBOW). Those models represent features that use the vector number. Synonym words can be found by using the cosine similarity function between the two vectors [12].

Cosine similarity is a statistical technique used to measure the similarity between two documents (d_1,d_2) represented by numeric vectors in the projection space. A cosine similarity value closer to one suggests similar documents; alternatively, a value that is closer to zero suggests dissimilar ones. Cosine similarity is calculated as shown in (2) [13].

$$Cos(d_1, d_2) = \frac{d_1, d_2}{(d_1.d_2)^{1/2} + (d_1.d_2)^{1/2}} \quad (2)$$

Data classification is the process of creating machine learning models in which a relationship exists between the features and classes of a dataset. Popular data classification algorithms are SVM [14], J48 [15] and NB [16]. SVM is a classification algorithm designed for binary-class problems. SVM classifiers create a decision boundary in a hyperplane that divides the data into two classes in the feature space using a non-probabilistic binary based on a linear function. The function determines a decision boundary that maximizes the margin between the support vectors. However, functions defining the decision boundary can be polynomial and radial based. The advantage of the SVM classifier is that it does not cause an overfitting problem from the model memorizing too many of the training set. Therefore, the model cannot classify the test dataset to its best ability [14]. In comparison, J48 is a Decision Tree classification algorithm. J48 classifiers select the feature with the highest information gain value, which is then used as the root node of the tree. The model is created using a top-down greedy search that selects features from the root node. The J48 classifier is suitable for large datasets because of its lower runtime [15]. Finally, NB classifiers use a conditional probability calculation. P (A | B) is the conditional probability or probability that event B occurs first and is followed by event A. P (A ∩ B) is the joint probability or the probability that event A and event B will both occur. P (B) is the probability that event B will occur. The NB classifier makes it easy to train models using a dataset with a large number of features, such as text datasets. The conditional probability calculation is shown in (3) [16].

$$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{B})}{\mathrm{P}(\mathrm{B})} \quad (3)$$

Performance measurements are the measurements of text classification models that assess their accuracy. However, this process may sometimes end up revising the model and evaluating the text mining process until the model is the most accurate. Accuracy is calculated from the correct classification of the model that considers all classes divided by all data, as shown in (4) [17].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

F-measure is an overall value that measures the correlation between precision and recall values, as shown in (5) [18].

$$Precision = TP/(TP + FP), Recall = TP/(TP + FN)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

AUC is the area under the receiver operating characteristic (ROC) curve graph. AUC is the area under the 2D graph to the x-axis (representing the FP) and the y-axis (representing the TP), as shown in (6) [19].

$$AUC = \frac{1+TP-FP}{2} \quad (6)$$

The Kappa coefficient is a statistic used to examine the consistency of the results of classification between two classes. The dataset used in the experiment does not have to have a normal distribution or non-parametric statistics. P_o is the observed probability of agreement, and P_e is the hypothetical expected probability of agreement, as shown in (7) [20].

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (7)$$

MCC is a measure of the efficiency classification results that is used with two-class datasets. The MCC value determines the balance of classification results with a value between -1 and +1 being calculated using TP, TN, FP and FN, as shown in (8) [21, 22].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

Runtime performance is calculated from the 3 components of the actual working time: train time, test time and model time [23].

2. PROPOSED ALGORITHM

The BWF algorithm was a domain-based feature selection technique performed using the BoW dataset and Word2Vec. This algorithm filtered the features of the BoW dataset to produce a new dataset for the creation of a text classification model. The advantage of this algorithm was that it created a BWF dataset smaller than the BoW dataset. The BWF algorithm included two steps. The first step involved creating the BoW dataset, consisting of the set of an instance where a *bow* such that each *bow* was instance 1 to instance *n*, as shown in (9).

$$BoW = \{bow_1, bow_2, \dots, bow_n\} \quad (9)$$

W was a set of features in the BoW dataset where *W* contained the set of features starting from feature 1 to feature *w*, as shown in (10).

$$W = \{feature_1, feature_2, \dots, feature_w\} \quad (10)$$

The second step involved a domain-based feature selection technique, performed using BoW and Word2Vec. The domain-based feature selection technique used three steps:

- Word2Vec was used to produce a pre-trained word embedding model from the methamphetamine tweet dataset. We used the Skip-gram model, an algorithm that generated the pre-trained word embedding model using Word2Vec. Tomas Mikolov suggested this algorithm, which was superior for infrequent words. Those words consisted of technical terms, slang name and synonym name. The Skip-gram model selected infrequent words to calculate the vector number. Thus, infrequent words had a higher-quality vector number than when using CBOW [12]. The pre-trained word embedding model consisted of 100-dimensional features represented by vector number. We defined the 100-dimensional features in focusing on runtime competencies that were used to create the pre-trained word embedding model from a large corpus.
- The set of domain-based features (SDBF) was created by measuring the cosine similarity between domain keywords in the pre-trained word embedding model. Our research used the keyword "*methamphetamine*" as the common name of methamphetamine.

$$SDBF = Cosine\ Similarity\ (Pre-trained\ word\ embedding\ model, "keywords")$$

The SDBF was sorted by descending cosine similarity. If the cosine similarity was equal to or greater than 0.8, those features were selected for inclusion as filter features of the BoW dataset. The SDBF contained the set of features starting from feature 1 to feature *w'*, as shown in (11).

$$SDBF = \{feature_1, feature_2, \dots, feature_{w'}\} \quad (11)$$

The BoW dataset was filtered to keep only the features in the SDBF. Next, the BoW dataset was considered based on the summed frequency in each instance of the dataset. If the sum frequency of an

instance was equal to zero, that instance was deleted from the BoW dataset. This research used the R programming package to implement the BWF algorithm [24]. The proposed data preprocessing technique consisted of the BWF algorithm, as shown in Figure 1.

Input: $\boldsymbol{BoW} = \{bow_1, bow_2, \dots, bow_n\}$
Methamphetamine Tweet Dataset
Output: $BWF\ dataset = \{bow'_1, bow'_2, \dots, bow'_m\}$.
Pre-trained word embedding model* = *Word2Vec(*Methamphetamine Tweet Dataset*)
SDBF* = *Cosine Similarity* (*Pre-trained word embedding model*, *"keywords")
SDBF is sorted by descending cosine similarity
For each $feature_i$ in ***SDBF***
  If cosine similarity value less than 0.8
…….. $feature_i$ is removed from ***SDBF***
  End If
End For
Return *SDBF* dataset ▷ $SDBF = \{feature_1, feature_2, \dots, feature_{w'}\}$
BWF dataset = Copy of ***BoW***
BWF dataset is **INNER Join** $\boldsymbol{W}$ and ***SDBF*** ▷ $W = \{feature_1, feature_2, \dots, feature_w\}$
For each bow'_i in ***BWFdataset***
  $|\boldsymbol{bow'_i}|$ is the sum frequency of an instance in $\boldsymbol{bow'_i}$
  If $|\boldsymbol{bow'_i}|$ equal to 0
…….. $\boldsymbol{bow'_i}$ is removed from ***BWF*** dataset
  End If
End For
Return *BWF* dataset

Figure 1. BWF algorithm

From Figure 1, the result of the BWF algorithm was a new dataset, called the "BWF dataset", which used the same text representation outcomes from the BoW dataset. This dataset was used for text classification in that the word frequency was used for the feature of the training with the classifier algorithm. However, the BWF dataset had fewer features and instances than the BoW dataset. The BWF contained the set of vectorization (*bow'*), where each vectorization was from instance 1 to instance *m*, as shown in (12):

$$BWF\ dataset = \{bow'_1, bow'_2, \dots, bow'_m\} \quad (12)$$

Proof:

Let *w* be the number of features in *BoW*. Let *SDBF* be the set of features. *SDBF* derives from the cosine similarity using the threshold of 0.8. Let *w'* be the number of features in *SDBF*. The *BWF* dataset is derived from *BoW* with only the features in *SDBF*. Thus, the number of features in the *BWF* dataset must be at most *w'*. Moreover, the *BWF* dataset is produced by removing (instance of) *BoW* in which the sums of all feature frequencies are equal to 0. Therefore, the number of instances in the *BWF* dataset must be less than that of *BoW*.

3. RESEARCH METHOD

This research consisted of two objectives. The first was the development of the "BWF" dataset. The second was the development of the TMTA, which consisted of the following steps: tweet collection, data preprocessing, classification, performance testing and hypothesis testing, as shown in the overview of the research framework in Figure 2.

3.1. Tweet collection

3.1.1. Synonym identification

This procedure involved the identification of keywords related to methamphetamine consisting of the common name, slang name and street name. These were collected and identified by the UK police [25]. In addition, we used the common name of methamphetamine to measure cosine similarity with Google News vectors [26] to look for additional slang names that had not been collected and identified by the UK police.

3.1.2. Tweet retrieval

Tweet retrieval is the selection of short text on Twitter related to methamphetamine that was posted by users in Southeast Asia, specifically Thailand, Indonesia, and Myanmar.

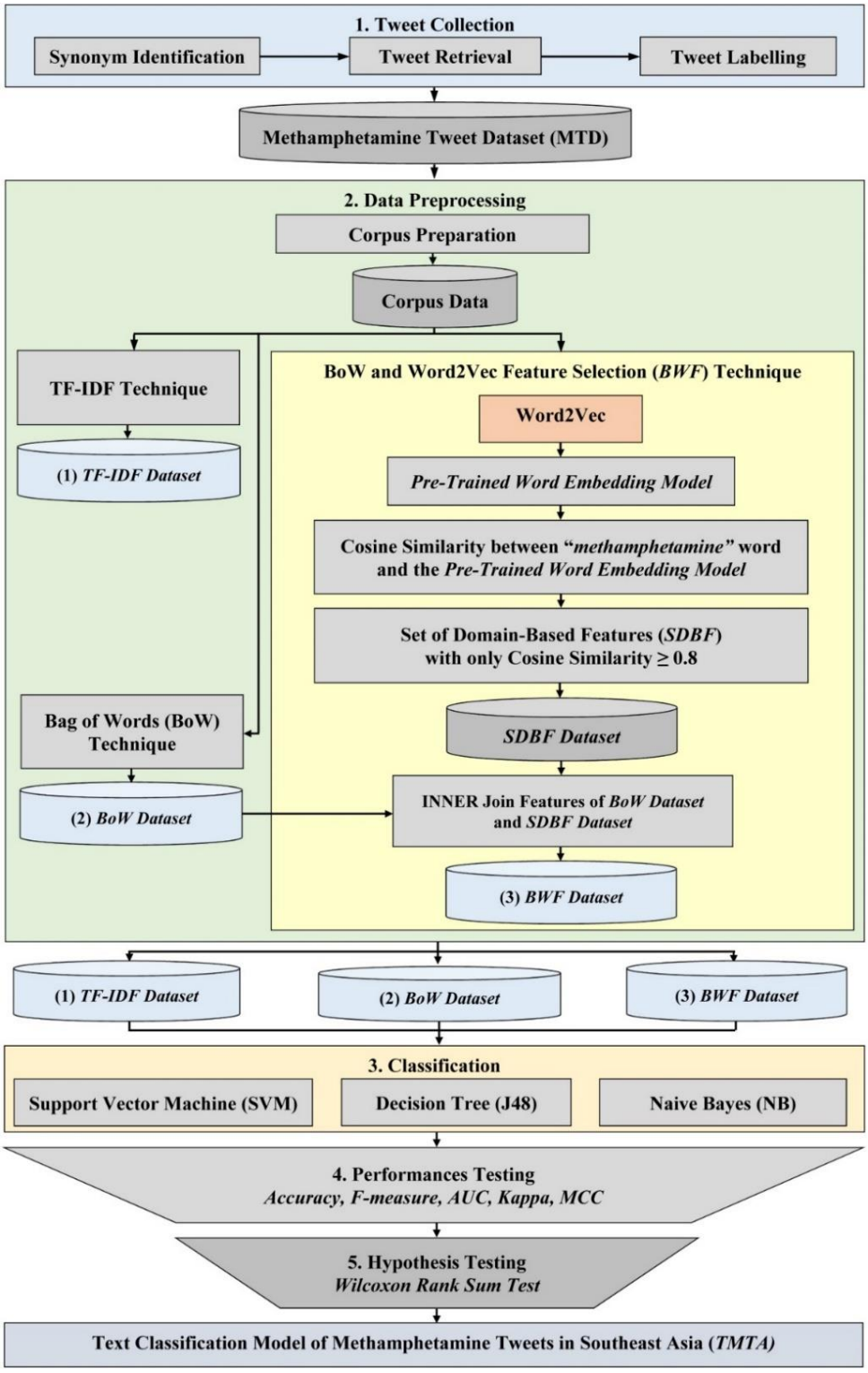

Figure 2. Research framework

3.1.3. Tweet labeling

Tweets were labeled by an expert from the Royal Thai Police Forensics Office into 2 classes: Non-abuse or abuse. Non-abuse tweets mentioned the penalty for using methamphetamine or its use as a medicine. The abuse class contained tweets about the illegal use of methamphetamine, including tweets promoting the use of methamphetamine, such as encouraging substance abuse to reduce obesity.

3.1.4. Methamphetamine tweet dataset (MTD)

We collected 2,899 tweets from online social media related to methamphetamine in Southeast Asia that an expert from the Royal Thai Police Forensics Office subsequently labeled. These data were divided into two classes: 2,170 instances of non-abuse and 729 instances of abuse, for a total of 23,175 words. The output of this step was MTD, whose properties are shown in Table 1.

Table 1. Characteristics of MTD

Instances		2,899
Number of Classes		2
Number of Class Members	Non-Abuse	2,170
	Abuse	729
Total Features (words)		23,175

3.2. Data preprocessing

This process consisted of corpus preparation, text representation and BWF.

3.2.1. Corpus preparation

Corpus preparation included stop word elimination and stemming. Stop word elimination involved removing some words that were not important and did not need to be further analyzed. Stop word elimination consisted of making all words lowercase, cutting markers, cutting tabs, cutting stop points and cutting stop words, such as “on”, “in”, “to” and “the”. Stemming was the modification of words that had the same stem meaning but were written differently, such as “eat” and “eating”. Stemming reduced the number of features of the methamphetamine dataset [27].

3.2.2. Text representation

The process of text representation was a part of NLP that converted text to vector. Vectorization created a set of vectors number representing text tweets that were used to create a text classification model. The classifier could operate on the text vectors. We used data preprocessing techniques consisting of BoW, TF–IDF and BWF, using BoW, a popular text vectorization model, as a baseline. If words appeared in the tweets, then the frequency was counted as 1; otherwise, it was counted as 0 [3, 4]. The TF-IDF algorithm, a data preprocessing technique that replaced the text with weight values, calculated the weight of importance that words used as a feature for each tweet. We determined that an important feature should not appear in every tweet. The TF–IDF method is widely used in text mining research [10, 11], while BWF represents the new data preprocessing technique that our research proposed. This algorithm performed the domain features selection of the BoW dataset.

3.3. Classification

Classification was the process of creating text classification models. In this study, the classification algorithms SVM [14], J48 [15] and NB [16], classifiers found in the Weka software, were used to create the text classification models. The Weka version 3.9 program, which is open source and widely used in research for this purpose, was used to develop the text classification models [28, 29].

3.4. Performances testing

We used 10-fold cross-validation for the measurement of TMTA performance using various metrics: accuracy [17], F-measure [18], AUC [19], Kappa [20], MCC [21, 22] and runtime [23]. The 10-fold cross-validation technique is a popular method to obtain reliable test results because all data points are used for training and validation; each data point is used to be tested exactly once [28].

3.5. Hypothesis testing

The Wilcoxon Rank Sum Test was used to investigate 5 different performance measurements (accuracy, F-measure, AUC, Kappa, MCC) between the proposed and candidate models to determine the differences in 5 performance measurements at a significance level of 0.05 [30, 31].

4. RESULTS AND DISCUSSION

This section describes and discusses the experimental results. It includes four sub-chapters, presented according to the two objectives and based on the characteristics of the BWF dataset, information gain, classification performance and hypothesis testing.

4.1. Characteristics of BWF dataset

The feature reduction performance using the BWF algorithm was compared with two popular techniques: BoW and TF–IDF. As Table 2 shows, the BWF dataset had fewer features (969) and instances (2,446) than the BoW and the TF-IDF datasets. The BWF algorithm was highly efficient at feature reduction. The experimental results demonstrated that the BWF dataset included 969 features out of the total 23,175 features in the methamphetamine tweet dataset. Table 2 shows the BWF dataset, which had a smaller number of features and instances than the BoW and TF-IDF datasets. Those features were filtered features of the BoW dataset using SDBF. Therefore, the BWF algorithm was effective at handling the semantic words associated with methamphetamine, such as slang names or synonyms for methamphetamine. This implementation was different from the BoW, as the latter reduces features by removing infrequent words.

Table 2. Comparison of data preprocessing techniques

Characteristic	Data Preprocessing Technique		
	BoW	TF-IDF	BWF
Number of Features	10,926	10,464	969
Number of Instances	2,899	2,899	2,446

4.2. Information gain

Information gain was applied to measure the quality of the features used to create a Decision Tree. The information gain tests for the BWF dataset identified several important features, including “meth”, “lab”, “crystal”, “ice”, “smoke”, “police”, “news”, “report”, “sexy” and “fat”. The words “meth” and “lab” were important features in the BWF dataset as they were used in tweets that mentioned laboratory-produced methamphetamine. The words “crystal” and “ice” are slang names for methamphetamine; both had high information gain, indicating the features’ potential for the prediction classes using the Decision Tree. Ten important features are shown in Table 3.

Table 3 shows the experimental results of the information gain that was used to test the feature quality of the BWF dataset. High information gain indicated the important features for the prediction classes based on the Decision Tree. Those features had strong power in classifying the classes based on the Decision Tree. Information gain showed important features such as “news”, “police” and “report” in the non-abuse class tweets; in contrast, “fat” and “sexy” were features of the abuse class tweets.

Table 3. Important features of BWF dataset using information gain

Ranked Feature	Information Gain (descending order)
meth	0.09514
lab	0.03701
crystal	0.03511
ice	0.02542
smoke	0.02329
police	0.02143
news	0.01640
report	0.01065
sexy	0.01048
fat	0.00555

4.3. Classification performance

The classification performance comparison of the three preprocessing techniques used to produce BoW, TF–IDF and BWF datasets are shown in Tables 4, 5 and 6. First, the performance of the SVM classifier with the BoW dataset had the highest accuracy (0.813), F-measure (0.803) and MCC (0.465). However, this classifier used with the BWF dataset had the highest AUC (0.720) and Kappa (0.461). Moreover, the BWF dataset had the lowest runtime (0.820 seconds) with the SVM classifier. Table 4 displays the classification performance comparisons of the three preprocessing techniques combined with SVM.

The decision tree using the J48 classifier with the BWF dataset had the highest scores in all measures, including accuracy (0.815), F-measure (0.818), AUC (0.763), Kappa (0.528) and MCC (0.529), and the lowest runtime (3.480 seconds). Table 5 presents the classification performance comparisons for this classifier. The NB classifier with the BoW dataset had the highest accuracy (0.795), F-measure (0.789), Kappa (0.428) and MCC (0.430). However, when combined with the BWF dataset, this classifier had the highest AUC (0.819) and the lowest runtime (0.400 seconds). The classification performance comparisons are shown in Table 6.

Table 4. Classification performance comparison using SVM

Measurement	SVM Classifier		
	BoW Dataset (baseline)	TF-IDF Dataset	BWF Dataset
Accuracy	0.813	0.812	0.805
F-measure	0.803	0.794	0.800
AUC	0.708	0.684	0.720
Kappa	0.456	0.424	0.461
MCC	0.465	0.446	0.463
Runtime (seconds)	33.310	11.580	0.820

Based on the classification performance comparisons in Tables 4, 5 and 6, the proposed model that combined the J48 classifier with the BWF dataset showed the best performance for the TMTA based on the four measures of accuracy, F-measure, Kappa and MCC. In comparison, the SVM classifier with the BWF dataset was the best based on runtime, and the NB classifier with the BWF dataset provided the highest AUC.

Table 5. Classification performance comparison using J48

Measurement	J48 Classifier		
	BoW Dataset (baseline)	TF-IDF Dataset	BWF Dataset
Accuracy	0.807	0.807	0.815
F-measure	0.804	0.805	0.818
AUC	0.723	0.735	0.763
Kappa	0.474	0.474	0.528
MCC	0.475	0.475	0.529
Runtime (seconds)	61.060	62.550	3.480

The results from Tables 4, 5 and 6 compare the performance measurements for SVM, J48 and NB, revealing that the model built on the J48 classifier and using the BWF dataset was the best. In short, this model provided the best performance measurements (accuracy, F-measure, Kappa, MCC). The highest accuracy was shown in terms of the correctness of the data classification using this model. The BWF dataset included 1,827 instances of non-abuse tweets and 619 instances of abuse tweets. This model could be predicted to correct 1,565 non-abuse tweets and 428 abuse tweets. Additionally, this model provided the highest F-measure values. This result showed that the model demonstrated accurate classification of the interest class, which was the abuse tweets.

Table 6. Classification performance comparison using NB

Measurement	NB Classifier		
	BoW Dataset (baseline)	TF-IDF Dataset	BWF Dataset
Accuracy	0.795	0.490	0.794
F-measure	0.789	0.495	0.785
AUC	0.797	0.762	0.819
Kappa	0.428	0.165	0.414
MCC	0.430	0.260	0.419
Runtime (seconds)	6.190	7.870	0.400

The AUC values of J48 with the BWF dataset were close to 1 as shown in Table 5, indicating that the classification results of J48 with the BWF dataset had high true positive values. The findings revealed that J48 with the BWF dataset highly classified the abuse class (here, an invitation tweet to consume methamphetamine). Table 5 shows the model generated using J48 with the BWF dataset, which had the highest Kappa and MCC values, suggesting high consistency in classification between the two classes (abuse or non-abuse).

The BWF dataset was fitted to the J48 classifier because the features in the BWF dataset were similar to the keyword "methamphetamine". Table 3 shows the features that had high information gain. Therefore, those features were used as a condition for classification based on the Decision Tree, and then the J48 classifier was used as a subset of the Decision Tree.

4.4. Hypothesis testing

As depicted in Table 7 the Wilcoxon rank sum test results suggested that the proposed model based on the J48 classifier using the BWF dataset was the best. This model was presented as TMTA because the five performance measurements (accuracy, F-measure, AUC, Kappa, MCC) were significantly higher than

for the six-candidate models with a P-Value of 0.043. However, J48 with the BWF dataset yielded performance measurements that were not significantly higher than NB using the BOW and BWF dataset with a P-Value of 0.225. The Wilcoxon rank sum test results for the performance measurements are shown in Table 7.

Table 7. Wilcoxon rank sum test for performance measurements

Proposed Model	Candidate Model	P-Value
J48 with BWF	SVM with TF-IDF	0.043
	SVM with BOW	
	SVM with BWF	
	J48 with TF-IDF	
	J48 with BOW	
	NB with TF-IDF	
	NB with BOW	0.225
	NB with BWF	

Table 7 shows the results of the Wilcoxon rank sum test, which was tested at a significance level of 0.05. The measured values for the accuracy, F-measure, AUC, Kappa and MCC of the proposed model were compared with the eight candidate models. The experimental results suggested these five performance measurements of the proposed model were better than for the six candidate models at a significance level of 0.05 with a statistical confidence level of 95 percent.

Therefore, the J48 classifier using the BWF dataset was used in developing the TMTA because this model provided the highest four performance measurements (accuracy, F-measure, Kappa and MCC) and provided a low runtime as shown in Table 5. Furthermore, this model provided significantly higher performance measurements than the six-candidate models as shown in Table 7.

Previous research created text classification models using tweet data based on SVM, J48 and NB classifiers. Although SVM with TF–IDF is still widely used for the development of text classification models [6-9], we found that the TMTA, using J48 with the BWF dataset, provided higher values for performance measurements than SVM with TF–IDF. In particular, the TMTA using J48 with the BWF dataset had a lower runtime than such widely used techniques as BoW and TF–IDF.

5. CONCLUSION

We proposed a new model, called the TMTA, to identify whether a Twitter tweet was related to methamphetamine use or abuse based on data extracted from Twitter in Southeast Asia. A vital process in the TMTA is data preprocessing. This research addressed the weakness of BoW in terms of feature selection using the BoW dataset and Word2Vec. A novel data preprocessing technique, the BWF algorithm, used the text vectorization method in the same way as the BoW dataset; however, the proposed BWF algorithm was applied using the feature selection of the BoW dataset to produce a BWF dataset. This approach resulted in a smaller number of features than such widely used techniques as BoW and the TF-IDF datasets. The new dataset was used for the TMTA dataset. The development of the TMTA consisted of four steps. First, we collected data with keywords related to methamphetamine from the Twitter data stream. Second, data preprocessing techniques were applied, including corpus preparation and text representation consisting of BoW, TF-IDF and BWF. Third, we experimented and proposed a text classification model using three candidate classifiers: SVM, J48 and NB. Lastly, we compared the performance of the various text classification models that were created from the above three classifiers using three data preprocessing techniques. The performance measurements included accuracy, F-measure, AUC, Kappa, MCC and runtime. Additionally, the TMTA model development used the J48 classifier with the BWF dataset. This model produced the highest values for accuracy (0.815), F-measure (0.818), Kappa (0.528) and MCC (0.529), high AUC (0.763) and low runtime (3.480 seconds) using the J48 classifier. These results showed that the proposed TMTA was fitted to the Twitter dataset collected in this study. The TMTA using J48 with the BWF dataset provided higher performance measurements than such traditional techniques as SVM with TF–IDF. Consequently, the TMTA using the J48 classifier could be converted to an if-then rule-based decision tree. This rule might be implemented for prototype software to help the police of the narcotics control board identify short messages related to drug abuse.

The BWF algorithm can be used for data preparation stemming from the development of a text classification model based on a different domain, such as amphetamine use in Thailand or illegal advertisements for nutritional supplements. Police have found tens of thousands of amphetamine networks on

social media that have cozened young juveniles into becoming members for distributing amphetamine. These networks offered promotions that were paid after transacting drugs. Furthermore, illegal advertisements for nutritional supplements are a problem in Thailand and have been widely sold using social media in that country. Most products (e.g. sexual enhancement products for men) exaggerate their properties. Both problems might be addressed to new investigations in future.

REFERENCES

[1] United Nations Office on Drugs and Crime, UNODC, "World Drug Report 2017," Vienna, Austria: United Nations Publication, 2017.

[2] N. Phan, S. A. Chun, M. Bhole and J. Geller, "Enabling Real-Time Drug Abuse Detection in Tweets," in *Proceeding of 33rd International Conference on Data Engineering (ICDE 2017)*, 2017, pp. 1510–1514.

[3] X. Sun, Y. Xiao, H. Wang and W. Wang, "On Conceptual Labeling of A Bag of Words," in *The 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, pp. 1326–1332.

[4] O. J. Ying, M. M. A. Zabidi, N. Ramli and U. U. Sheikh, "Sentiment Analysis of Informal Malay Tweets with Deep Learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 2, pp. 212–220, Jun. 2020.

[5] B. T. Johns and R. K. Jamieson, "A Large-Scale Analysis of Variance in Written Language," *Cognitive Science*, vol. 42, no. 4, pp. 1360–1374, May 2018.

[6] J. R. Ragini and P. R. Anand, "An Empirical Analysis and Classification of Crisis Related Tweets," in *Proceeding of International Conference on Computational Intelligence and Computing Research (ICCIC2016)*, 2016, pp. 1–4.

[7] Y. Wang, Z. Zhou, S. Jin, D. Liu and M. Lu, "Comparisons and Selections of Features and Classifiers for Short Text Classification," *IOP Conference Series: Materials Science and Engineering*, vol. 261, 2017, pp. 1–7.

[8] S. Ghosh, P. K. Srijith, and M. S. Desarkar, "Using Social Media for Classifying Actionable Insights in Disaster Scenario," *International Journal of Advances in Engineering Sciences and Applied Mathematics*, vol. 9, no. 4, pp. 224–237, Dec. 2017.

[9] G. Burel and H. Alani, "Crisis Event Extraction Service (CREES)-Automatic Detection and Classification of Crisis-related Content on Social Media," in *Proceeding of 15th International Conference on Information Systems for Crisis Response and Management*, 2018.

[10] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018.

[11] I. Yahav, O. Shehory, and D. Schwartz, "Comments Mining with TF-IDF: The Inherent Bias and Its Removal," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 437–450, 2018.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Proceeding of Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] M. Vijaymeena and K. Kavitha, "A Survey on Similarity Measures in Text Mining," *Machine Learning and Applications: An International Journa*l, vol. 3, no. 2, pp. 19–28, 2016.

[14] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics-Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[15] P. Kapoor, and R. Rani, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning," *International Journal of Engineering Research and General Science*, vol. 3, no. 3, pp. 1613–1621, 2015.

[16] S. Wang, L. Jiang, and C. Li, "Adapting Naive Bayes Tree for Text Classification," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 77–89, 2015.

[17] A. J. Larner, "New Unitary Metrics for Dementia Test Accuracy Studies," *Progress in Neurology and Psychiatry*, vol. 23, no. 3, pp. 21–25, 2019.

[18] D. Hand and P. Christen, "A Note on Using the F-measure for Evaluating Record Linkage Algorithms," *Statistics and Computing*, vol. 28, no. 3, pp. 539–547, 2018.

[19] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PloS one*, vol. 10, no. 3, pp. 1–21, 2015.

[20] S. Mishra and Nitika, "Understanding the Calculation of the Kappa Statistic: A Measure of Inter-Observer Reliability," *International Journal of Academic Medicine*, vol. 2, no. 2, pp. 217–217, Sep. 2016.

[21] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric," *PloS One*, vol. 12, no. 6, pp. 1-17, Jun. 2017.

[22] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[23] T. Doan and J. Kalita, "Predicting Run Time of Classification Algorithms Using Meta-Learning," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 6, pp. 1929–1943, Jul. 2016.

[24] R Core Teams, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: http://www.R-project.org/

[25] Telegraph Media Group, "Police given 3,000 Word 'a to z of Drugs Slang' to Stay Ahead of Criminals," 2019. Available: http://www.telegraph.co.uk/news/uknews/law-and-order/6519172/Police-given- 3000-word-A-to-Z-of-drugs-slang-to-stay-ahead-of-criminals.html

[26] Google, "Word2Vec-GoogleNews-Vectors," 2016, [Online], Available: https://github.com/mmihaltz/Word2Vec-GoogleNews-Vector

[27] N. Seman and N. A. Razmi, "Machine Learning-Based Technique for Big Data Sentiment Extraction," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 3, pp. 473–479, Sep. 2020.
[28] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, "The WEKA Workbench, Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques," *Fourth Edition, Morgan Kaufmann*, 2016.
[29] S. Hussain, N. A. Dahan, F. M. Ba-Alwib and N. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 9, no. 2, pp. 447–459, Feb. 2018.
[30] R. S. M. de Barros, J. I. G. Hidalgo and D. R. de Lima Cabral, "Wilcoxon Rank Sum Test Drift Detector," *Neurocomputing*, vol. 275, pp. 1954–1963, Jan. 2018.
[31] A. K. Dwivedi, I. Mallawaarachchi and L. A. Alvarado, "Analysis of Small Sample Size Studies Using Nonparametric Bootstrap Test with Pooled Resampling Method," *Statistics in medicine*, vol. 36, no. 14, pp. 2187–2205, Mar. 2017.

BIOGRAPHIES OF AUTHORS

Narongsak Chayangkoon received his B.Sc. degree in Information Science from Walailak University, Thailand, in 2002 and his M.Sc. degree in Computer Technology from King Mongkut's University of Technology North Bangkok, Thailand in 2007. He is currently a Ph.D. candidate in the Department of Computer Science, Kasetsart University, Bangkok, Thailand. He is interested in data preprocessing techniques, text classification and machine learning.

Anongnart Srivihok received her B.Sc. degree in Microbiology from Chulalongkorn University, Thailand in 1978, her M.Sc. degree in Computer Science from the University of Mississippi, USA in 1984, and her Ph.D. degree in Information Systems from the Central Queensland University, Australia in 1998, respectively. She is presently Associate Professor of Computer Science, Kasetsart University, Bangkok, Thailand. She is interested in data mining, machine learning and knowledge management.