http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 2, December 2017: 53 - 64
e-ISSN: 2289-2192

A COMPARATIVE STUDY OF THE ENSEMBLE AND BASE CLASSIFIERS PERFORMANCE IN MALAY TEXT CATEGORIZATION

HAMOOD ALI ALSHALABI
SABRINA TIUN
NAZLIA OMAR

ABSTRACT

Automatic text categorization (ATC) has attracted the attention of the research community over the last decade as it frees organizations from the need of manually organized documents. The ensemble techniques, which combine the results of a number of individually trained base classifiers, always improve classification performance better than base classifiers. This paper intends to compare the effectiveness of ensemble with that of base classifiers for Malay text classification. Two feature selection methods (the Gini Index (GI) and Chi-square) with the ensemble methods are applied to examine Malay text classification, with the intention to efficiently integrate base classifiers algorithms into a more accurate classification procedure. Two types of ensemble methods, namely the voting combination and meta-classifier combination, are evaluated. A wide range of comparative experiments are conducted to assess classified Malay dataset. The applied experiments reveal that meta-classifier ensemble framework performed better than the best individual classifiers on the tested datasets.

Keywords: Feature selection. individual classifiers. ensemble classifiers. Malay text classification

INTRODUCTION

Text Categorization is defined as the way of making a decision if a certain piece of text belongs to one of sets of prescribed categories. As a significant stage in the Natural Language Processing system, it is convenient in indexing and later restoring texts. Moreover, Text Categorization is advantageous for content analysis, and a lot of other roles (Lewis & Gale 1994). On the other hand, a crucial problem may be arisen during data mining and, therefore, Machin Learning ML comes from the big confluence of information in the Internet due to the increase electronic documents, and information libraries available (Mitchell 1999).

The idea of Text Categorization is to specify one document to one or more categories, depending on its contents. In this regard, the automatic text categorization process foresees a set of tasks universally recognized by the research community (Abdullah et al 2005). These tasks include features design in which the corpus processing, extraction of relevant information, feature selection and feature weighting processes are performed. In addition, these tasks include training in which a machine learning classifier is trained using a set of labelled documents. The last task is the task of testing in which the classifier accuracy is evaluated through the use of a set of pre-labelled documents (i.e. test-set) which are not used in the training phase.

The key idea behind combining individual classifiers is that every individual classifier's certain strengths and weaknesses are emerged accordingly. Hence, it could be argued that they can benefit from the strengths of individual classifiers and their weaknesses could be positively enhanced. In addition, classifiers are combined in order to make use of their strengths. Therefore, combined methods are becoming more popular as they allow to overcome

the weaknesses of single supervised approaches. Classifiers can be composed to be multiple distinct classifiers by selecting the best classifier to be used in different situations or contexts (Srinivas et al. 2009). In this paper, Combined Classifiers have been investigated along with the performance of different methods for classifiers combination.

The most of the work in this area were carried out for the English text and other well-studied languages. Up to date, there are very few and scarce works have been carried out for the Malay, which differ morphologically and syntactically from other languages, due to the lack of resources for managing Malay Text Classification (MTC). Consequently, the need to construct the resources and tools for MTC is a growing. This motivates us to apply an appropriate methods for Malay Text which has different morphologically to can achieve the best results.

The paper is organized as follows: Section 1 has been devoted for brief Introduction and Section 2 sheds some light on the Methodology used. The different key techniques and approaches are described in Section 3, reviewing related works in Text Categorization. Section 4 is allocated for presenting the experiment setup and discussing the experimental results. Sections 5 is specified for the study conclusion focusing on the realized findings. Finally, Section 6 recommends future trends in this subject matter.

METHODOLOGY

The methodology used in the present paper, Malay Text Classification, is shown in Figure 1. First, tasks of pre-processing were used to eliminate the incomplete and inconsistent data. The purpose beyond this process was to perform further data mining functionality. Secondly, Feature selection methods were carried out to discriminate terms for training and classification. Thirdly, k-NN, NB, and N-gram, are applied on Malay ATC. The k-NN, NB, and N-gram methods are used due to their simplicity and effectiveness and their accurateness Fourthly, Combination algorithm will be used to select the best result from the three results obtained from the three single classifiers, k-NN, NB, and N-gram, on Malay text. Finally, shows evaluation method for measuring the correctness of our finding.

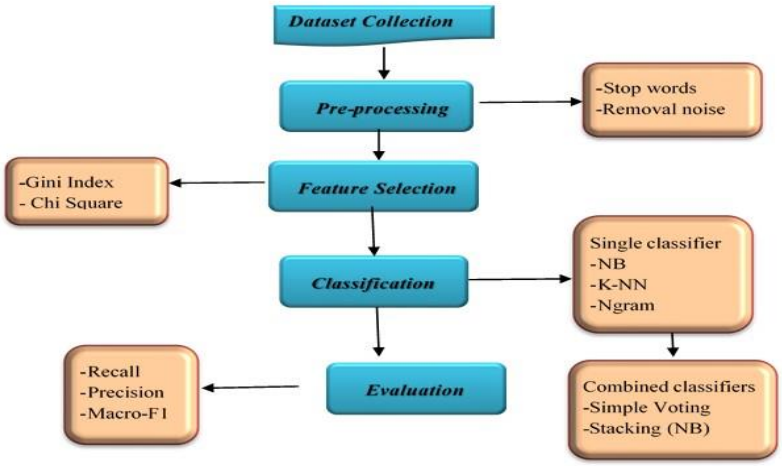

FIGURE 1. Illustration of the methodology of Malay Text Classification

PRE-PROCESSING

In order to evaluate the used classification algorithms, several experiments were conducted. The performance of these classification algorithms was measured to classify the Malay corpus

used in (Alshalabi et al, 2013) study. The corpus would be divided into six categories namely: *Business, Crime, History, Health, Religion and Sports.*

Before indexing all of the documents, including training sets and test sets, they were all passed the preprocessing phase. The phase was beneficial because it worked on minimizing the index size, raised accuracy and merged categorization activities. However, not all words of a document seemed to be significantly equivalent to their meanings . Some words have more meaning compared to others. Therefore, it was crucial to pre-process the text in the dataset collection to identify the proper words to be employed as features. Each specific word appears in a document was defined as a feature. In the preprocessing step, advantageous text operations could be performed such as removing stop words and noise removal (Baeza-Yates&Ribeiro-Neto 1999). Further description of these two operations has been described in detail in the next sub sections. *Case folding* is the phase of changing uppercase to lowercase in the document, then, the elimination of punctuation other than the "a" to "z" letter which is considered as the delimiter character.

Tokenizing and Noise Removal is the phase of splitting sentence to words. With the word's splitting first, the string that has been input will be simpler. Therefore, in each word, the string is shown according to the space which split it with that form. In this way, changing process a word stem becomes easier. On the other hand, *Noise Removal* is the process of refining words and removing special characters, numbers, and symbols which add up to the noise in the training dataset.

Stop-Words Malay language has a large number of stop-words (i.e. words having little content-bearings). The highly frequent words existed in documents collection, considered as noisy in the text, (e.g. pronouns, prepositions, conjunctions, etc.) are called stop-word. Malay words such as, *apabila, bagi, dalam, para,* and *untuk* are considered as stop words as shown by (Ahmad 1995) .The stop words are removed since they do not convey any important information, and thus will reduce the text representation and improve the performance of the classification. Conversely, the words that are more relevant to each document will be left.

FEATURE SELECTION (FS)

Feature Selection (FS) method is one of the most crucial tasks that improves the performance of text classification due to the selection of the most predictive features. In other words, FS develops the performance of text classification tasks in terms of learning speed and effectiveness and also reduces the number of data dimensions. Moreover, FS removes irrelevant, redundant, and noisy data (Sebastiani 2002). In this section, further explanation will be provided on the feature selection methods where Gini Index (GI) and Chi Squire are used in our Malay ATC:

Gini Index (GI):

$$\text{Gini}(t) = \sum_{i=1}^{|C|} p(t|c_i)^2 \, p(t|c_i)^2 \qquad (1)$$

A novel GI algorithm is introduced by Shang et al (2007) based on the Gini-Index theory. The researchers constructed a new measure function of Gini index. They consider feature t's condition probability, combining posterior probability and condition probability as the whole measure function to depress the affection when the class is unbalanced. The main idea of Gini index is that, first, removing the situation that feature words do not appear, second, introducing concentration between classes and within-class dispersion to the traditional information gain feature selection method.

Chi Square **(CS)** measures the absence of independence between t (term) and c (category) (Rogati&Yang 2002; Yang&Pedersen 1997) . It can be calculated as follows:

$$\chi^2(c,t) = \frac{N \times (\text{AD-BC})}{(A+C)(B+C)(A+B)(C+D)} \quad (2)$$

$$\chi^2_{max}(t) = max_i(\chi^2(t,c_i) \quad (3)$$

where A is the number of documents that contain the term, t, and also belong to category, c. B is the number of documents that contain the term, t, but do not belong to category, c. C is the number of documents that do not contain the term, t, but belong to category, c. D is the number of documents that do not contain the term, t, and do not belong to category, c. N is the number of training documents (Thabtah et al. 2009).

CLASSIFICATION METHODS

As has been mentioned earlier, three classifier methods are selected and used in Malay Text Classification. Single Classifiers Methods as: (k-NN, NB and N-gram methods) and Classifier Combination (Simple Voting and Stacking combination) due to their simplicity, effectiveness and accurateness methods are assumed, as follows:

SINGLE CLASSIFIER METHODS

K-NEAREST NEIGHBOR (K-NN)

The k-NN is a well-known example-based classifier. It is one of the most popular classification techniques due to its simplicity and accuracy. The k-NN is also known as lazy learner, since it delays the decision on how to generalize beyond the training data until each new query instance is encountered. In order to categorize a document, the k-NN classifier organizes scores of the document's neighbours among the training documents. Then, it uses the class labels of the k most similar neighbours. Given a test document d, the system finds the K nearest neighbours among training documents. The similarity score of each nearest neighbour document to the test document is used. The weighted sum in k-NN classification is written as follows:

$$Score(d_i,d) = \sum_{d_j=KNN\,(d)} sim(d,d_j)\cdot\partial(d_j,c_j) \quad (4)$$

Where KNN (d) indicates the set of *K* nearest neighbours of a document d. If d_j belongs toc_i, then $\delta(d_j,c_i)$ equals 1, or other-wise 0. For test document d, it should belong to the class that has the highest resulting weighted sum. In order to compute $sim\,(d,d_j)$, the Euclidean distance is used representing the usual manner in which humans think of distance in the real world (He et al. 2000):

$$D_{\text{Euclidean}}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (5)$$

NAIVE BAYES (NB)

The NB algorithm is widely used as an algorithm for document classification. It is a probability-based classifier. Based on the features, independent probability value is calculated for each and every model. NB is often used in text category tasks based on Bayes' formula:

$$P(C_i|d) = \frac{P(C_i)P(d|C_i)}{P(d)} \quad (6)$$

Where $P(C_i|d)$, is the posterior probability of class C_i given a new document d $P(C_i)$ is the probability of class C_i which can be calculated by:

$$P(C_i)\frac{N_i}{N} \quad (7)$$

Where N_i, is the number of documents assigned to class C_i and N is the number of classes, $P(d|C_i)$ is the probability of a document d given a class C_i, and $P(d)$ is the probability of document d, and because of the independence assumption of NB, the probability of document d can be calculated by:

$$P(C_i|d)P(C_i)\prod_{k=1}^{n} p(t_k|C_i) \quad (8)$$

Where t_k is a feature that occurs with class C_i, and also we can calculate $p(t_k|C_i)$ by:

$$P(t_k|C_i)\ \frac{1+n_{ki}}{1+\sum_{h=1}^{l} n_{hk}} \quad (9)$$

Where n_{ki} is the total number of documents that contain feature t_k and belong to class C_i. The number 'l' indicates the total number of distinct features in all training documents that belong to class Ci. NB calculates posterior probability for each class, and then assigns document d to the highest posterior probability's class, i.e.

$$C(d) = \underset{i=}{\operatorname{argmax}}\,|C|\ (P(C_i|d)) \quad (10)$$

To explain how Naïve Bayes model works, two classes have been assumed. The first class is economic and the second is not-economic. Four training documents are gained, three of them are from economic class and the last one is not from the economic class, as shown in Table 1.

TABLE 1. Data for parameter estimation in NB classifier examples

	N	words in document	in c= economic ?
Training set	1	ekonomi <u>selangor</u> ekonomi	economic
	2	ekonomi ekonomi <u>melaka</u>	economic
	3	ekonomi <u>penang</u>	economic
	4	<u>kedah</u> polis ekonomi	Not economic
Test set	5	ekonomi ekonomi ekonomi <u>kedah</u> polis	?

Given a test document, the multinomial parameters are needed to classify the test document and they are considered as the priors $p(c) = \frac{3}{4}$ and $p(\bar{c}) = \frac{1}{4}$ following conditional probabilities:

$$p(\text{ekonomi }|\text{economic }) = \frac{(5+1)}{8+6} = \frac{3}{7}$$

$$p(\text{polis }|\text{economic }) = p(\text{Kedah}|\text{economic }) = \frac{(0+1)}{8+6} = \frac{1}{14}$$

$$p(\text{ekonomi }|\text{not economic }) = \frac{(1+1)}{3+6} = \frac{2}{9}$$

$$p(\text{polis }|\text{not economic }) = p(\text{Kedah }|\text{not economic }) = \frac{(1+1)}{3+6} = \frac{2}{9}$$

The denominators are (8+6) and (3+6) because the lengths of all documents in class economic are 8 and the length of all documents is not in class economic 3, respectively. The size of the distinct terms is 6 as the vocabulary consists of six terms.

$$p(\text{c} = \text{economic }\ |\text{d5 }) = \frac{3}{4} * \left(\frac{3}{7}\right)^3 \left(\frac{1}{14}\right)\left(\frac{1}{14}\right) = 0.0003$$

$$p(\text{c} = \text{not economic }\ |\text{d5 }) = \frac{1}{4} * \left(\frac{2}{9}\right)^3 \left(\frac{2}{9}\right)\left(\frac{2}{9}\right) = 0.0001$$

Where as $p(\text{c} = \text{economic }\ |\text{d5 }) > p(\text{c} = \text{not economic }\ |\text{d5 })$, then the document d5 is in class economic

N-GRAM CLASSIFIER

An N-gram is a continuous sequence of n characters or n words of a longer portion of a text (Mohan et al, 2010). This research paper intends to use the character level N-grams classifier. In the N-gram training process, the N-gram profile needs to be generated. The generated N-gram profile consists of the text which is spilt into tokens consisting of letters only. The most frequent N-grams are the ones kept. This gives us the N-gram profile for the document. For the purpose of classifying each documents, each document needs to go through the text preprocessing phase, then, the N-gram profile is generated as described above (Ogada, 2016). The N-gram profile of each document will then be compared with the profiles of all documents in the training classes (class profile) in terms of similarity. Specifically, the cosine similarity measurement is used to measure the similarity between two documents, the training document Di and test document D_j:

$$Sim_{cosine}(D_i, D_j) = \frac{\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sqrt{\sum_{k=1}^{m} W_{ik}^2 \times \sum_{k=1}^{m} W_{jk}^2}} \quad (11)$$

CLASSIFIER COMBINATION

In this stage, an ensemble (classifier combination) approach is applied for the sake of selecting results based on the output of the three classifiers. The selection algorithm is employed as the main task in this methodology to determine the accuracy of the combined classifiers via choosing the best answer out of a set of three answers. Here, we list the selection algorithms used in our Malay TC

MAJORITY (SIMPLE VOTING)

In the simple voting mechanism each base classifier model has a single vote. For each test document, this vote is given to the class label returned by the base model. After all base classifiers are voted, the class label having maximum votes is selected as the correct class label for that document. The class that appears as the choice of the largest number of classifiers is picked as the answer. If all classifiers disagree, the algorithm will choose the result of the tagger with highest accuracy. In the (simple voting) each classifier has a single vote. (Srinivas et al. 2009). To explain how the voting algorithm work on TC, suppose that we have three classifiers as in our case (classifier 1 (S1), classifier 2 (S2) and classifier 3 (S3) and we have two classes one is economic and the second class is not economic. Let us assume that we want to assign test document x to either class economic or class not economic. As shown in Table 2. , the final decision depends on the majority. If two or more classifiers agree that the document is economic, then the final decision is economic, and two or more classifiers agree that the document is not economic, then the final decision is economic.

TABLE 2 example demonstrate the Voting Combination

S1 decision	S2 decision	S3 decision	(voting) decision
economic	economic	economic	economic
economic	economic	not economic	economic
economic	not economic	economic	economic
economic	not economic	not economic	not economic
not economic	economic	economic	economic
not economic	economic	not economic	not economic
not economic	not economic	economic	not economic
not economic	not economic	not economic	not economic

STACKING COMBINATION)

The stacking combination consists of two phases. In the first phase, a set of base-level classifiers is generated. In the second phase, a meta-level classifier is learnt combining the outputs of the base-level classifiers (Xia et al, 2011). When using a meta-classifier for combination, the outputs of all the labels of the class of the participating classifiers are used as features for meta-learning (Koprinska et al, 2007 ; Mitchell, 1997). In our case, to combine the output of the three classifiers Naïve Bayes, k-NN and N-gram decision, we use as meta-classifier the Naïve Bayes. The formula (12) of the NB as meta-classifier, given the output of three classifiers, $R_{1,}R_2, R_3$:

$$P(C_i|R_1, R_2, R_3) = \frac{P(C_i)P(R_1R_2R_3|C_i)}{P(R_1R_2R_3)} \quad (12)$$

Where $P(C_i|R_{1,}R_2, R_3)$ is the posterior probability of class C_i given the new output of the three classifiers $R_1R_{,2}$, R_3, $P(C_i)$ is the probability of class C_i .

EVALUATION

All algorithms are evaluated using a 5-fold cross-validation measurement tool. To measure the performance of these classification methods, we use the Macro-averaged (Macro-F1) measure. This measure combines Recall and Precision in the following way

$$\text{Precision} = \frac{TP}{PT + FP} \quad (13)$$

$$\text{Recall} = \frac{TP}{PT + FN} \quad (14)$$

$$F1 = \frac{2Pr \times Re}{Pr + Rp} \quad (15)$$

$$F_1^{macro} = \frac{1}{m}\sum_{i=1}^{m} F_1(i) \quad (16)$$

In which, True Positive (TP) is the set of document that is correctly assigned to the given category. False Positive (FP) is the set of documents that are incorrectly assigned to the category. False Negative (FN) is the set of documents that is not assigned incorrectly to the category. On the other hand, True Negative (TN) is the set of documents correctly not assigned to the category. To explain how to evaluate the classification algorithms work, let us assume that we have a set of documents as given for matches of human and computer document assignments in table 3

RELATED WORK

Feature ranking and selection are essential parts of text classification, and a lot of methods and approaches have been investigated and applied to feature selection for text classification. Most of FS methods can be classified into two groups; information theory ranking methods such as chi-square and mutual information, and information retrieval ranking methods such as document frequency and odd ratio (Ghareb et al. 2014). For example, Yang and Pedersen (1997) and Thabtah (2007) evaluated five methods of feature selection namely: DF, chi square $\chi 2$, term strength (TS), information gain (IG), and mutual information (MI) with K-NN. The

Naïve Bayesian classifier as presented by Chen et al. (2009). They pointed that these methods perform better than other feature selection approaches when they are experimented with English and Chinese text collections.

Chiang et al. (2008) modified TF-IDF and utilized it with ARM and category priority to construct their classifiers. Based on Alshalabi et al. (2013), this study depends on NB, N-gram and k-NN classifiers methods with the two feature selection methods, Chi and GI of the TC in order to enhance Malay TC. The first experiment examined the overall performance of the NB, N-gram and k-NN classifiers with the two feature selection methods, Chi and GI, are applied to reduce the dimension of feature spaces on Malay TC. According to Sanwaliya et al. (2010), NB and KNN classifiers are considered as a single classifier and their accuracy have been investigated using Reuters 21578 corpus data. Then, these classifiers are combined according to a proposed method (NB-KNN). The obtained results have shown that accuracy I significantly improved. In Nejat et al. (2012), NB, SVM, and DT classifiers are considered as single classifier and their accuracy have been investigated using corpus data. Then, these classifiers are combined according to a proposed Meta classifier (Boosting, Voting, and Bagging). The results have shown that the comparison between base and ensemble classifiers in terms of the best values for accuracy show that NB ensemble classifier is considered as a better alternative, although their classifications' accuracy turns to be equal. In Srinivas et al. (2009), the classifier combination methods and concept-based dimensionality reduction techniques are used for robust and scalable text classification.

The experimental evaluation confirms the hypothesis that combination based meta-classifiers give better accuracy than individual classifiers for a popular textual dataset, the Reuters 21578 news dataset. Additionally, text classification methods were first proposed in the 1950s where the word frequency was used to classify documents automatically. Applications of machine learning techniques help reduce the manual effort required for analysis and the accuracy of the systems also improved through the use of these techniques. Interestingly, several text mining software packages are available in the market. In addition, many machine learning methods have been proposed for text categorization in previous years including N-gram (Suzuki et al. 2012; Farhoodi et al. 2011), Naïve Bayes (Mccallum&Nigam 1998; Fan et al. 2001), and k-nearest neighbor (Hua&Sun 2001).

Zhang et al. (2015) provided a study devoted for character-level convolutional networks for text classification. They compared a large number of traditional and deep learning models using several largescale datasets. The, analysis showed that character level convent is an effective method. Additionally, the model of comparisons depends on many factors, such as dataset size, if the texts are curated, and choice of alphabet. The study by Johnson and Zhang (2016) viewed that the model is considered as a special case of a general framework which jointly trains a linear model with a non-linear feature generator consisting of 'text region embedding + pooling'. In their study, the authors discovered a more sophisticated region embedding method using Long Short-Term Memory (LSTM). LSTM can embed text regions of variable or possibly large sizes. (Relatively, et al, 1998) introduced support vector machines for TC. The study provides both theoretical and empirical evidences that SVMs are significantly suitable for TC.

EXPERIMENTAL EVALUATION

This section is concerned with dividing the data into two subgroups. The first subgroup is called the single classifiers or individual classifiers. The second subgroup is called combined classifiers. For the purpose of this work, two kinds of experiments are carried out.

SINGLE CLASSIFIERS RESULTS

The first set of experiments show the performance of the individual based classifiers. In order to test the efficiency of the three classifiers k-NN, NB and N-gram, with the two feature reduction methods on Malay text Categorization, these methods are evaluated individually and features are selected from feature space at different size: 100, 200, 300, 400, 500 and 600. The results are presented in terms of macro-averaged F-measure where the averaged values are calculated across the whole 5-fold cross-validation experiments. The overall performance of the NB, N-gram and k-NN classifiers with the two feature selection methods, Chi-square and GI applied to reduce the dimension of feature spaces, has been examined precisely.

TABLE 3: The performance (Macro-F1) of single (feature selection methods vs. features sizes).

	k-NN		**NB**		**N-gram**	
	Chi	GI	Chi	GI	Chi	GI
100	89.14%	89.53%	90.92%	90.92%	78.35%	78.84%
200	88.39%	90.38%	93.73%	93.21%	77.37%	78.70%
300	90.72%	90.72%	92.82%	94.66%	79.01%	79.01%
400	89.24%	90.92%	93.30%	94.12%	78.67%	78.76%
500	89.03%	89.12%	93.45%	94.52%	78.82%	78.51%
600	86.13%	87.00%	94.00%	94.13%	78.40%	78.40%

At this phase, the effects of the individual feature selection method on classifiers performances have also been examined. The results of the performance (see Table 5) is displayed with features ranked in a degrading order and feature space at different sizes: 100, 200, 300, 400, 500 and 600. In Table 3, the best performance of 94.66 is the NB classifier when 300 of the features selected using Chi-square feature selection. In addition, the best accuracy of 90.72 with k-NN classifier is achieved when 300 of the features selected by GI method are used, and the highest performance with N-gram classifier has been obtained when 300 of the features by GI method are used. When the classifier performances are compared, the NB algorithm achieves a higher performance than that of the k-NN and N-gram algorithms. Thus, it is obvious that the highest performance is obtained when the feature selection operations are made by GI. This observation indicates that the k-NN and NB classifiers are both suitable for Malay Text Categorization.

In order to examine the overall performance based on document categories, all of the parameters for the three classifiers, k-NN, NB and N-gram, are fixed according to their best results in Table 3. The experimental results with the k-NN, NB and N-gram for Malay. As seen in Fig. 2, the NB achieves the best result in Sports, *Business, Crime,* and *History* domains while the NB obtains its best result in *Sport* and *Business* domains.

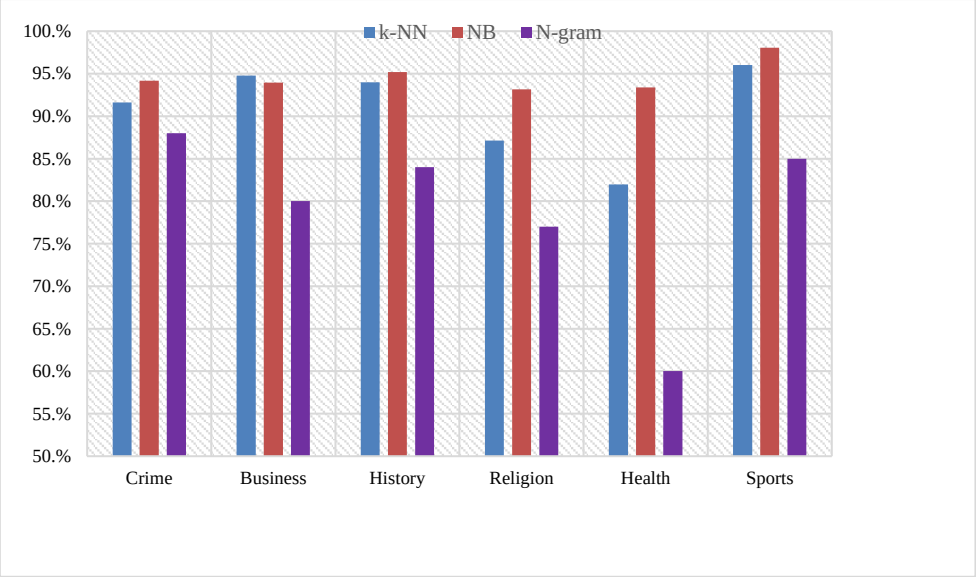

FIGURE 2. The performance (F-measure) on each class of single classifiers

CLASSIFIER COMBINATION RESULTS

The second set of experiments has been combined into the three classifiers values which examine the classifier combination. This methodology is to determine the accuracy of the combined classifiers by choosing the best answer giving a set of three answers. There are two types of classifier combinations namely: Voting Combination and Stacking Combination. Through these experiments, the following are realized: The best performance of Voting Combination reaching 95.84%, is achieved when 500 of the features are selected by GI method. On the contrary, the worst performance of Voting Combination, being 92.14%, is achieved when 100 of the features are selected by Chi squire method. In Table 4, the best performance of Stacking Combination is 94.39% achieved when 300 of the features are selected by GI method while the lowest performance of Stacking Combination, reaching 91.23%, is achieved when 400 of the features are selected by Chi squire method. It is clear that higher performance is obtained when the FS operations are made by GI. The results obtained are convergent. It is further obvious that the results achieved by the Stacking Combination algorithm are better than that those scored by individual classifiers. However, the Voting combination achieves better results compared to Stacking Combination.

TABLE 3. The performance (Macro-F1) of Meta classifier (feature selection methods vs. features sizes).

	Voting		Stacking (Zhang, #47)	
	Chi	GI	Chi	GI
100	92.77%	92.14%	91.44%	91.96%
200	95.29%	94.34%	93.06%	93.71%
300	95.55%	94.75%	92.94%	94.39%
400	95.15%	95.71%	91.23%	92.83%
500	94.61%	95.84%	92.56%	93.63%
600	95.54%	95.57%	93.10%	94.00%

The experimental results of the Voting and Stacking Classifiers for Malay Text Categorization are shown in Fig 3. The NB achieves the best result in Sports, *History, Crime,* and *Business* domains whereas the Voting achieves its best result in *Sport* and *History* domains.

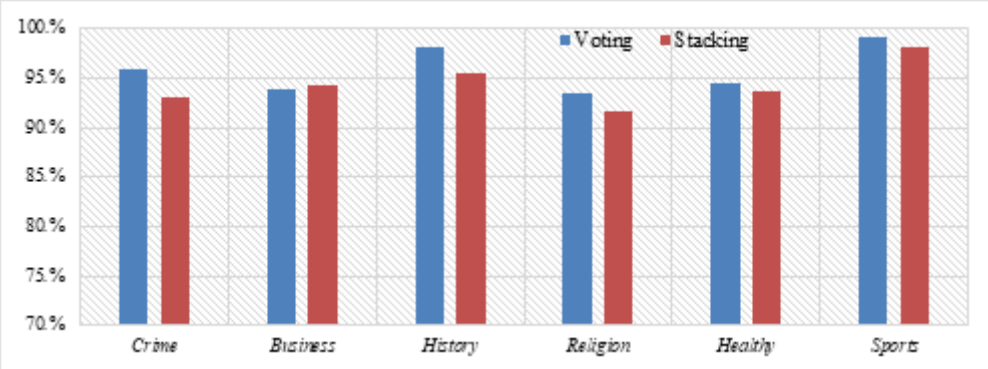

FIGURE 3. The performance (F-measure) on each class of Meta classifier

CONCLUSION

In conclusion, the three ML methods namely NB, N-gram, and k-NN with the two FS methods, Chi and GI, are applied to reduce the dimension of feature spaces on Malay TC. Moreover, two classifiers combination methods, namely, Voting combination and Stacking Combination methods have been evaluated. The results have shown that among the individual classifiers, the NB classifier with the two FS methods, Chi and GI, achieved the best performance in term of macro-F1. However, the results of Voting Combination method are higher than that those of NB. Moreover, the study findings further indicate that in Malay TC, the two features algorithms behave in the same way.

RECOMMENDS AND FUTURE WORKS

In this study, it could be reflected that experiments with a small dataset can show significant results. However, for future trends in this subject matter, it could be planned to expand the size of the corpus. Hence, we are preparing for collecting big dataset in future. This can provide a wider chances to carry out several experiments and improvements on the feature selection phase. The reason beyond that may due to the fact that many problems can be faced when proposing a method for Malay text Classification with satisfactory accuracy. For instance, the wide range of Malay language vocabulary, creates challenges such as the lack of text representation and the lack of important words identification. This problem arises from the misleading words that should be removed at the beginning of the performance stage.

REFRENCE

Ahmad, F. 1995. A Malay language document retrieval system: an experimental approach and analysis. *Universiti Kebangsaan Malaysia, Bangi*.

Abdullah, M. T. & Ahmad, F. 2005. Improvement of Malay Information Retrieval Using Local Stop Words. pula 630(0.31.

Alshalabi, H., S. Tiun, N. Omar & M. Albared 2013. Experiments on the use of feature selection and machine learning methods in automatic malay text categorization. *Procedia Technology* 11: 748-754.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. Modern Information Retrieval. ACM press New York.

Fan, Y., Zheng, C., Wang, Q., Cai, Q. & Liu, J. 2001. Using Naive Bayes to Coordinate the Classification of Web Pages. Journal of software 12(9): 1386-1392.

Farhoodi, M., Yari, A. & Sayah, A. 2011. N-Gram Based Text Classification for Persian Newspaper Corpus. Digital Content, Multimedia Technology and its Applications (IDCTA), 2011 7th International Conference on, hlm. 55-59.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant

features. *Machine learning: ECML-98*: 137-142.
Johnson, R. & T. Zhang 2016. Supervised and semi-supervised text categorization using one-hot LSTM for region embeddings. *stat* 1050: 7.
He, J., Tan, A.-H. & Tan, C.-L. 2000. A Comparative Study on Chinese Text Categorization Methods. Proceedings of PRICAI'2000 International Workshop on Text and Web Mining, hlm. p24-35.
Hua, S. & Sun, Z. 2001. Support Vector Machine Approach for Protein Subcellular Localization Prediction. Bioinformatics 17(8): 721-728.
Ghareb, A. S., A. R. Hamdan & A. A. Bakar 2014. Integrating noun-based feature ranking and selection methods with Arabic text associative classification approach. *Arabian Journal for Science and Engineering* 39(11): 7807-7822.
Lewis, D. D. & Gale, W. A. 1994. A Sequential Algorithm for Training Text Classifiers. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, hlm. 3-12.
Mccallum, A. & Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 workshop on learning for text categorization, hlm. 41-48.
Mitchell, T. M. 1999. Machine Learning and Data Mining. Communications of the ACM 42(11): 30-36.
Nejat, M. H., Aghazarian, V. & Hedayati, A. R. Comparative Study of the Performance of Ensemble and Base Classifiers in Text Data Categorization.
Rogati, M. & Yang, Y. 2002. High-Performing Feature Selection for Text Classification. Proceedings of the eleventh international conference on Information and knowledge management, hlm. 659-661.
Sanwaliya, A., Shanker, K. & Misra, S. C. 2010. Categorization of News Articles: A Model Based on Discriminative Term Extraction Method. Advances in Databases Knowledge and Data Applications (DBKDA), 2010 Second International Conference on, hlm. 149-154.
Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. ACM computing surveys (CSUR) 34(1): 1-47.
Shang, Wenqian, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang. "A novel feature selection algorithm for text categorization." *Expert Systems with Applications* 33, no. 1 (2007): 1-5.
Srinivas, M., Supreethi, K. & Prasad, E. 2009. Combining the Classifiers and Lsi Method for Efficient and Accurate Text Classification. Journal of Information Technology and Knowledge Management 2(2): 263-267.
Suzuki, M., Yamagishi, N. & Tsai, Y.-C. 2012. Chinese Text Categorization Using the Character N-Gram. Information Theory and its Applications (ISITA), 2012 International Symposium on, hlm. 722-726.
Thabtah, F., Eljinini, M., Zamzeer, M. & Hadi, W. 2009. Naïve Bayesian Based on Chi Square to Categorize Arabic Data. proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, hlm. 930-935.
Wenqian, S., Houkuan, H., Yuling, L., Yongmin, L., Youli, Q. & Hongbin, D. 2006. Research on the Algorithm of Feature Selection Based on Gini Index for Text Categorization [J]. Journal of Computer Research and Development 10(001.
Yang, Y. & Pedersen, J. O. 1997. A Comparative Study on Feature Selection in Text Categorization. MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, hlm. 412-420.
Zhang, X., J. Zhao & Y. LeCun 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*. pp. 649-657.

Hamood Ali Alshalabi
Sabrina Tiun
Nazlia Omar
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Selangor.

Received: 18 July 2017
Accepted: 31 December 2017