JAIST Repository
https://dspace.jaist.ac.jp/

Title	Factors affecting sentiment prediction of malay news headlines using machine learning approaches
Author(s)	Alfred, Rayner; Wong, Wei Yee; Lim, Yuto; Obit, Joe Henry
Citation	Communications in Computer and Information Science, 652: 289-299
Issue Date	2016-09-18
Type	Journal Article
Text version	author
URL	http://hdl.handle.net/10119/14737
Rights	This is the author-created version of Springer, Rayner Alfred, Wong Wei Yee, Yuto Lim, Joe Henry Obit, Communications in Computer and Information Science, 652, 2016, 289-299. The original publication is available at www.springerlink.com, http://dx.doi.org/10.1007/978-981-10-2777-2_26
Description	Second International Conference, SCDS 2016, Kuala Lumpur, Malaysia, September 21-22, 2016, Proceedings

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Factors Affecting Sentiment Prediction of Malay News Headlines Using Machine Learning Approaches

Rayner Alfred[1], Wong Wei Yee[1], Yuto Lim[2], and Joe Henry Obit[1]

[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Sabah, Malaysia
{ralfred,joehenry}@ums.edu.my, huiyee0529@hotmail.com,
[2]School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Japan
ylim@jaist.ac.jp

Abstract. Malay language is a major language that is in used by citizens of Malaysia, Indonesia, Singapore and Brunei. As the language is widely used, there are abundant of text articles written in Malay language that are available on the internet. This has resulted in the increasing of the Malay articles published online and the number of articles has increased greatly over the years. Automatically labeling Malay text articles is crucial in managing these articles. Due to lack of resources and tools used to perform the topic selection automatically for Malay text articles, this paper studies the factors that influence the performances of the algorithms that can be applied to perform a topic selection automatically for Malay articles. This is done by comparing the contents of the articles with the corresponding topics and all Malay articles will be assigned to the appropriate topics depending on the results of the classification process. In this paper, all Malay articles will be classified by using the *k*-Nearest Neighbors (*k*-NN) and Naïve Bayes classifiers. Both classifiers are used to classify and assign a topic to these Malay articles according to a predefined set of topics. The effectiveness of classifying these Malay articles using the *k*-NN classifier is highly dependent on the distance methods used and the number of Nearest Neighbors, *k*. Thus, this paper also assesses the effects of using different distance methods (e.g., Cosine Similarity and the Euclidean Distance) and varying the number of clusters, *k*. Other than that, the effects of utilizing the stemming process on the performance of the classifiers are also studied. Based on the results obtained, the proposed approach shows that the *k*-NN classifier performs better than the Naïve Bayes classifier in classifying the Malay articles into their respective topics. In addition to that, the stemming process also improves the overall performances of both classifiers. Other findings include the application of Cosine Similarity as the distance measure has improved the performance of the *k*-NN classifier.

Keywords: Topic Selection, Feature Extraction**,** Classification, Clustering.

1 Introduction

Malay language is one of the languages that is widely spoken by people in the Southeast Asia especially in Malaysia, Singapore, Brunei, Indonesia and South of Thailand [18]. In fact, Malay language is the major language that is used by the citizens of Malaysia, Singapore and Brunei. As the language is widely used, there are abundant of text articles written Malay language and they are available on the internet [12]. The number of Malay articles published on the internet increases dramatically due to the advancement of internet-related technologies. As the number of these Malay articles or documents increases greatly, the efficiency of documents classification based on a predefined set of topics is becoming more important [19].

Topic selection is defined as the task of selecting the appropriate topic for an article based on the contents of the article. Topic selection has becoming as one of the impotant techniques in the studies as it enables documents to be assigned to a specific topic category from the set of predefined topics [5]. In other researches, topic selection is also known as topic identification [5], topic spotting [20], text categorization [9], or text classification [2]. A conventional topic selection framework consists of a pre-processing, feature extraction, feature selection and classification stages [5]. This paper studies the effectiveness of the topic selection methods using the *k*-NN and Naïve Bayes classifiers. The effectiveness of classifying these Malay articles using the *k*-NN classifier is highly dependent on the distance methods used and the number of Nearest Neighbors that is used, *k*. Thus, this paper also assesses the effects of using different distance methods (e.g., Cosine Similarity and the Euclidean Distance) and varying the values of k. Other than that, the effects of utilizing the stemming process on the performance of the classifiers are also studied.

The rest of this paper is organized as followed. Section 2 explains some of related works. Section 3 outlines the framework of the classification based topic selection for Malay articles. Section 4 describes the experimental setup and discusses the results obtained in investigating the effects of using different distance methods (e.g., Cosine Similarity and the Euclidean Distance) and varying the values of *k* on the performance of the *k*-NN classifier. The effects of utilizing the stemming process on the performance of the classifiers are also studied. Finally, Section 5 concludes this paper.

2 Related Works

There are researches conducted and models proposed that investigate methods to find the best and optimum way to perform the topic selection process for English articles [23]. However, there are not many works conducted that focus on the task of performing topic selection for Malay articles [3]. Thus, this has motivated this paper to propose a general framework of topic selection for Malay articles. Nevertheless, a study has been conducted to assess the performances of several machine learning methods coupled with several different features selection methods in performing the

topic selection for Malay articles [3]. Based on the results obtained, the *k*-Nearest Neighbor classifier performance is found to be better compared to the others classifier. However, the text preprocessing process had being neglected in the experiments done by Alshalabi. Text pre-processing is a vital process in any topic selection or text classification as it is the main task of extracting the features from the text documents. It is also known as the feature extraction process in topic selection as generally the features exist in the text documents are all words [1]. The importance of this process can be seen as many similar researches with tasks involving text documents include this process [11][1]. Text pre-processing is the first step in a topic extraction or topic selection process as it reduces the noise in the documents or texts by removing the unnecessary terms [13]. The following includes the three pre-processing tasks such as tokenization, stop words removal, punctuation removal, stemming and finally the construction of document-texts matrix that represents the Term-Frequency and Inverse Document Frequency (TF-IDF) [2].

Tokenization is one of the lexical analysis techniques. Tokenization is a process of forming tokens from an input stream of characters where a token is a string of one or more characters. Tokenization is the process of breaking a stream of text into tokens that can be words, sentences, phrases, symbols or other meaningful elements or terms to help contribute to the task that is under study [5]. Stop words removal is a process of removing the words that have little lexical meaning in the documents [5]. These words are unlikely to contribute to the distinctiveness of the texts and examples of the stop words are the conjunctions, pronouns and prepositions words. Punctuation removal is a process of removing all the punctuation that exists in the text during the text pre-processing. Punctuation is the usage of the special characters such as "/", "?", "!", ";" and many more. The uses of these special characters are to enhance and disambiguate the meaning of a sentence and those characters should be removed as it does not bring any lexical meaning which contributes to the analysis of topic selection to identify the topic of the text. Stemming refers to the transformation of a word from its root words to its stem. The stemming algorithm used to stem Malay text was different with those used to stem English text [15]. In order to stem Malay text, the suffixes, prefixes and infixes must be removed in a proper order [15]. As the Malay words are formed with a combination containing one or more syllable, thus is possible to perform stemming by analyzing the pattern of the syllable than analyzing the word character by character [6]. Lee's proposed stemmer consists of two components where first they perform the process of syllabification where it separate a Malay word into its syllables and the second component is the Malay stemming rules where it is used to identify the morphological structures of the Malay word to be stemmed. Term Frequency–Inverse Document Frequency (TF-IDF) is one of the terms weighting model that is used in the text processing to represent a document. It includes the use of the bag-of-words model, which is widely used in Information Retrieval (IR) and text mining [2]. TF-IDF is often used and also is the most popular to be used as the weighting scheme. Term Frequency, TF is the frequency that the one term t in a document d, represents the count of the particular term exists in the document. Inverse Frequency Document, IDF is the inverse document frequency of that term t across all the

documents in the document sets [16]. The IDF influences the weighting of the terms that exist in all the documents in the document sets as the IDF value for the terms that exists in all documents will be zero or less, resulting the TF-IDF value will become insignificant.

In determining the topic label of a single document, a text classification can be used to classify the documents into the appropriate topic. Text classification is a task of assigning a category to any single piece of text or document or articles. Thus, the text classification technique also is used to automate the topic selection as it allows the documents to be identified by a label, subject or topic. Text classification also allows the grouping of the documents by common topics. There are two most commonly used algorithms which is the *k*-Nearest Neighbors (*k*-NN) algorithm and the Naïve Bayes algorithm. The *k*-Nearest Neighbor algorithm, k-NN, is the most widely used in classification and clustering algorithm as it is simple, fast and effective [22]. It is also known as the lazy learning algorithm [10]. The *k*-NN algorithm is also applied as a classifier and *k*-NN classifier is also one of the favorite and widely used classifiers due to its ease in implementation yet powerful as it often shows good performance [21]. The *k*-NN classifier will classify the text documents accordingly the k number of neighbors. It classifies the text documents to the class or topic with the most votes which are determined based on the k nearest neighbors [4]. Naïve Bayes is a probabilistic classification that based on the Bayesian probability. It requires training data to learn the classes' probabilities. This classifier is called naïve as the Naïve Bayes classification is a classifier that assumes all the attributes are conditionally independent given the classes [17]. Despite that the naïve assumptions and its simplicity to be implemented, the Naïve Bayes classifier proved to be an effective classifier [7]. This classifier classifies a document, d in to a class, c by learning a probability value from the training data [14].

3 Topic Selection for Malay Articles

This section outlines the framework used to automate the topic selection for Malay articles.

3.1 Text Processing Tasks

Several text preprocessing tasks are described before classifying the articles into its respective topics. Three major tasks will be performed in this phase which is the tokenization, punctuation removal, Malay stop words removal and Malay words stemming. For the stop words removal, all the stop words found in the predefined list are removed. Example of stop words includes "*adalah*", "*ialah*", "*dan*" and many more. Malay words stemming is another text preprocessing task that will remove all the prefixes and suffixes from a Malay term or word so that the term will be transformed back into its root word. This paper will also investigate the effects of the stemming process on the topic selection for Malay articles. The following shows the list of prefix and suffix [6].

Table 1. List of Prefixes and Suffixes

Type	Substring
Prefix	*'ber', 'per', 'ter', 'mem', 'pem', 'meng', 'peng', 'men', 'pen', 'pe', 'me', 'be', 'ke', 'se', 'te', 'di'*
Suffix	*'nya', 'kan', 'an', 'i', 'kah', 'lah', 'pun', 'man', 'ku', 'mu'*

The following is the procedures of the Malay stemming algorithm that is used to perform the stemming process.

Step 1: Get the term from the term list.
Step 2: Check if the term is a root word by checking its number of letters where the threshold set is 5. If it is less than 5 letters, it is the root word, then exit the stemming process or else proceed.
Step 3: Check the first few letters against the Prefix list, if a match found, remove the prefix and add the appropriate letter in front if necessary.
Step 4: Check the term's letter threshold, if it is less than 5, exit or else proceed.
Step 5: Check the last few letters against the Suffix list, if a match found, remove the suffix.
Step 6: Check the term's letter threshold, if it is less than 5, exit or else proceed.
Step 7: Repeat the whole process again from Step 2 to remove multiple Prefix and Suffix on the term.

There are several more rules that will be needed as some prefixes will modify the first letter of the root word when the prefix is added to it. Thus, there is a set of rules for the first letter modification when applying the stemming as to restore the word back to its root word.

Rule 1: if the first letter is 'm', then change it to 'p' or 'f'.
Rule 2: if the first letter is 'n', then change it to 't'.
Rule 3: if the first letter is 'y', then change it to 's'.
Rule 4: if the first letter is vowel ('a', 'e', 'i', 'o', 'u'), then add 'k' at the front.

3.2 Term Weighting

A weight is assigned to each term by using the TF-IDF term weighting scheme. The weighting was done by applying the TF-IDF computation which is shown below.

$$TF - IDF = \left(1 + \log f_{i,j}\right) * \log \frac{N}{n_i} \quad (1)$$

where, $f_{i,j}$ refers to the frequency of the term t_i in the document d_j, N is the total number of the documents and n_i refers to the number of n documents that the term t_i exist in.

3.3 Identification of Topics Based on Classification

The topic selection for each article will be done by using the *k*-Nearest Neighbor and Naïve Bayes classifier. The *k*-NN classifier will classify the document into its topics or classes based on the classes of the *k* nearest neighbors of the document. The *k*-NN classifier uses the distance measure to compute the distance between documents. In this paper, the distance measure that will be used for comparison in their effectiveness is the Cosine similarity and the Euclidean distance methods. The Cosine Similarity is one of the methods used in order to compute the distance between two documents by computing the degree of similarity of the two documents. The Cosine similarity will produce a value where the closer the value to 1 means that both of the documents are similar to each other. The formula of the cosine similarity is shown below.

$$S_{dj,cp} = \sum_{dt \in N_k(d_j)} similarity\ (d_j, d_t) \times T(d_t, c_p) \tag{2}$$

where, $N_k(d_j)$ refers to the set of the *k* nearest neighbors in d_j in the training set, $similarity(d_j,d_t)$ refers to the Cosine formula for vector model and finally, $T(d_t,c_p)$ refers to the function that returns value of 1, if d_j is belongs to class c_p and 0 otherwise. The cosine formula is denoted as

$$sim(d_j, d_t) = \frac{d_j . d_t}{|d_j| \times |d_t|} \tag{3}$$

where, d_j is a vector of document j and d_t is a vector of training document *t*.

Euclidean distance is also another method which is commonly used to calculate the distance between two documents. The smaller value of Euclidean distance between two documents indicates that they are close to one and another. The Euclidean distance is computed as shown below.

$$Dist_{d_i,d_j} = \sqrt{\sum_{N=1}^{N} (x_{d_i} - x_{d_j})^2} \tag{4}$$

where, $Dist_{d_i,d_j}$ refers to the distance of d_i from d_j, N is number of terms or features in d_i and x_{d_i} and x_{d_j} are terms or features exists in the documents.

The Naïve Bayes will also be applied to classify the text documents by computing the probability of the document with respect to the class. It will assign the document to the class with the highest probability value computed [17].

The evaluation method used to evaluate the performance of each of the proposed approach is by applying the accuracy measure. It is a simply accuracy evaluation which computes the percentage of the documents whether it is correctly classified or not.

$$Accuracy = \frac{Number\ documents\ correctly\ assigned}{Total\ number\ of\ test\ documents} \times 100\% \tag{5}$$

4 Experimental Setup

This section describes the experiments that are carried out in this work. The experiments are outlined in order to achieve the following objectives,

a) To investigate the effectiveness of the applied stemming algorithm in reducing the number of terms or features extracted.
b) To investigate the best *k* value for the *k*-NN classifier.
c) To identify which classifier performs better in the proposed approach.

There are 1000 Malay news articles that will be used in this study. These news articles are retrieved from *Bernama* archive and *theStar* website. The news articles are retrieved and annotated manually in order to make sure that these documents are categorized into the appropriate topic labels. These topic labels include Financial, Politics, Sports, Entertainments and General. The following describes how the experiments are carried out.

Experiment 1: The dataset will not undergo the stemming process and the distance method used to compute the distance is the Cosine Similarity and the experiment is repeated with different values of *k* for *k*-NN classifier where $k = 1, 3, 5, 7$.

Experiment 2: Experiment 1 is repeated but using Euclidean distance.

Experiment 3: Repeat Experiment 1 by adding the stemming process into the pre-processing phase. Then, compute the distance by using the Cosine Similarity and repeat the experiment with different values of *k* for *k*-NN classifier.

Experiment 4: Repeat Experiment 3 by changing the distance computation with Euclidean distance computation.

Experiment 5: Repeat Experiments 1 and 3 by changing the Naïve Bayes classifier.

For the experiments, each of the experiment is validated by using the 2/3 fold cross validation method. The data sets are firstly divided into 3 different sets so that a different set of documents will be used as the training and test data while performing the cross validation.

Table 2. Prediction accuracies obtained for the *k*-NN classifier with different approaches

	Without Stemming								With Stemming							
	Cosine Similarity				Euclidean Distance				Cosine Similarity				Euclidean Distance			
k-NN's k value	1	3	5	7	1	3	5	7	1	3	5	7	1	3	5	7
CV1	97.3	**97.3**	96.7	97.3	95.3	96.0	**95.3**	95.3	97.3	**98.0**	96.7	96.7	95.3	96.0	**94.7**	94.7
CV2	92.0	**96.7**	94.7	94.0	84.7	90.0	**90.7**	88.7	92.7	**98.7**	96.7	94.7	83.3	93.3	**93.3**	92.0
CV3	96.0	**97.3**	96.0	95.3	94.0	94.7	**94.7**	89.3	96.7	**98.0**	97.3	96.7	93.3	94.7	**94.7**	89.3
Mean	95.1	**97.1**	95.8	95.5	91.3	93.6	**93.6**	91.1	95.6	**98.2**	96.9	96.0	90.6	94.7	**94.2**	92.0
Std. Dev	2.76	**0.35**	1.01	1.66	5.78	3.15	**2.50**	3.65	2.50	**0.40**	0.35	1.15	6.43	1.35	**0.81**	2.70

The results shows that *k*-NN classifier performs better when the stemming process is performed prior to the classification task, the Cosine similarity is used as the distance measure and when k = 3. The *k*-NN classifier has a higher accuracy with Cosine similarity as the distance measure compare to Euclidean distance as the distance measure. Thus, a conclusion can be drawn from the result in table 2 is that the *k*-NN classifier performs a better accuracy in classifying the documents when it uses cosine similarity as its distance measure, when its $k = 3$ and with the stemming algorithm.

On the other hand, the performance of the Naïve Bayes classifier also performs better with the stemming algorithm. The accuracy of the classifier is higher when it applied the stemming algorithm, 88.7%±3.46 compared to without stemming which is 83.1%±5.37. However, the result also shows that the *k*-NN classifier outperforms the Naïve Bayes classifier. The k-NN classifier has the accuracy above 90% for all the approaches while Naïve Bayes classifier only reaches 83.1% and 88.7% in both approaches, with and without performing the stemming process. This shows the *k*-NN classifier was indeed performed better than Naïve Bayes classifier in topic selection of Malay articles.

Table 3**.** Prediction accuracies obtained for the *k*-NN and Naive Bayes classifiers

	K-Nearest Neighbor Classifier				Naïve Bayes Classifier	
	Without Stemming		With Stemming			
	Cosine Similarity	Euclidean Distance	Cosine Similarity	Euclidean Distance	Without Stemming	With Stemming
CV1	97.2	95.5	97.2	95.2	80	86.7
CV2	94.4	88.5	95.7	90.5	89.3	92.7
CV3	96.2	93.2	97.2	93	80	86.7
Mean	**95.9**	92.4	**96.7**	92.9	83.1	**88.7**
Std. Dev	1.42	3.54	0.85	2.35	5.37	3.46

The results also show that the stemming does help improving the performances of the classifier. Performing the stemming process also improves the prediction classification by reducing the number of features or terms in each document.

5 Conclusion

In conclusion, the proposed approach for topic selection for Malay articles has been described in detailed in this paper. Based on the results obtained from the experiments carried out in this paper, the proposed approach to topic selection for Malay articles by performing only feature extraction does show promising results. The *k*-NN classifier is the best classifier to be used for topic selection of Malay articles. The *k*-NN classifier also performs better with Cosine Similarity as its distance method and the best *k* value for the *k*-NN classifier was also identified in the experiment where the best value for *k* was 3. The obtained prediction performance is found to be better when the stemming process is performed prior to the classification process as redundant features are eliminated that could decrease the prediction performance. This can be seen in the results obtained from the experiments where both *k*-NN classifier and Naïve Bayes classifiers have improvement in the accuracy when stemming algorithm was included. However, the stemming algorithm still required refinement and it can be further improve. As mentioned before, a suggestion on improving the stemming algorithm will be adding the use of root word dictionary which is a list of root words so that the modification of the words after removal of prefix can be done correctly and effectively.

References

1. Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee and Khairullah Khan: A Review of Machine Learning Algorithms for Text-Documents Classification. (2010)
2. Baeza-Yates, R. A., and Ribeiro-Neto, B. A.: Modern Information Retrieval (2nd edition.). Pearson Education Ltd. (2011)
3. Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, and Mohammed Albared: Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. (2013)
4. Harun Uguz: A Two-Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorithm. (2011)
5. J.D. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba and J.M. Montero: Topic Identification Techniques Applied to Dynamic Language Model Adaptation for Automatic Speech Recognition. (2014)
6. JunChoi Lee, Rosita Mohamad Othman and Nurul Zawiyah Mohamad: Syllable-based Malay Word Stemmer. (2013)
7. Liangxiao Jiang and Harry Zhang: Learning Instance Greedily Cloning Naïve Bayes for Ranking. (2005)
8. Mangalam, Sankupellay and Subbu Valliappan: Malay-Language Stemmer. (2006)
9. Manning, C. D., Raghavan, P., and Schütze, H.: Introduction to Information Retrieval. Cambridge University Press. (2008)
10. Meenakshi and Swati Singla: Review Paper on Text Categorization Techniques. (2015)
11. Nordianah Ab Samat, Masrah Azrifah Azmi Murad, Muhammad Taufik Abdullah, Rodziah Atan: Malay Documents Clustering Algorithm Based on Singular Value Decomposition. (2008)
12. Normaly Kamal Ismail, Nur Hamizah Mat Saad. Sidi Bukhari Sidi Omar, Tengku Mohammad Tengku Sembok: 2D Visualization of Terms and Documents in Malay Language. (2013)

13. Rim Koulali, Mahmoud El-Hajy and Abdelouafi Meziane: Arabic Topic Detection using Automatic Text Summarisation. (2013)
14. S. K. Thakur and V. K. Singh: A Lexicon Pool Augmented Naïve Bayes Classifier for Nepali Text. (2014)
15. Tengku Mohd T. Sembok, Zainab Abu Bakar and Fatimah Ahmad: Experiments in Malay Information Retrieval. (2011)
16. Wei Yong-qing, Liu Pei-yu and Zhu Zhen-fang: A Feature Selection Method based on Improved TFIDF. (2008)
17. Zhengchang Qin: Naïve Bayes Classification Given Probability Estimation Trees. (2006)
18. Mohd Yunus Sharum, Muhammad Taufik Abdullah, Md Nasir Sulaiman, Masrah Azrifah Azmi Murad, Zaitul Azma Zainon Hamzah: MALIM – A New Computational Approach of Malay Morphology. (2010)
19. Youngjoong, Ko and Jungyun Seo: Automatic Text Categorization by Unsupervised Learning. (2000)
20. Erik Wiener, Jan O. Pedersen, Andreas S. Weigend: A Neural Network Approach to Topic Spotting. (1995)
21. P. Viswanath and T. Hitendra Sarma: An Improvement to k-Nearest Neighbor Classifier. (2011)
22. Qu Chao, Yuan Ruifen, and Wei Xiaorui: KNNC: An Algorithm for K-Nearest Neighbor Clique Clustering. (2013)
23. J. Tanha, J. de Does and K. Depuydt: An LDA-based Topic Selection Approach to Language Model, Adaptation for Handwritten Text Recognition, Proceedings of Recent Advances in Natural Language Processing, pp. 646–653, Hissar, Bulgaria, Sep 7–9 2015.