TOPIC SELECTION FOR MALAY ARTICLES

LEOW JIA REN

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2015

TOPIC SELECTION FOR MALAY ARTICLES

LEOW JIA REN

SUBMITTED IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH

2015

CONFIRMATION

NAME	:	**LEOW JIA REN**
MATRIC NO.	:	**BK 11110142**
TITLE	:	**TOPIC SELECTION FOR MALAY ARTICLES**
DEGREE	:	**BACHELOR DEGREE OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)**
VIVA DATE	:	**29th JUNE 2015**

CERTIFIED BY

1. SUPERVISOR
Assc. Prof. Dr. Rayner Alfred

2. EXAMINER
Dr. Mohd Hanafi Ahmad Hijazi

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excepts, equations, summaries and references, which have been duly acknowledged.

22 Jun 2015 .....

Leow Jia Ren

BK 11110142

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and appreciation to my supervisor, Assoc. Prof. Dr Rayner Alfred for all his advices, guidance and support in this research work that leads to the completion of this thesis. He has been a very patience mentor in guiding me throughout the entire project. This thesis would have never been able to be completed in time and in a correct path without his guidance. I would like to thank Dr. Mohd Hanafi Ahmad Hijazi, my examiner for his valuable guidance, advices and recommendations in this research work that leads to the completion and enhancement in its quality. I would like to express my thanks to others lecturers, fellow coursemates, friends and family for their supports and helps throughout the whole process of completing this research whether were directly or indirectly. Lastly, I would like to thanks again to all the people that I mentioned above for their guidance, helps and supports and I wish them good luck for their future undertaking.

Leow Jia Ren
Jun 2015

ABSTRACT

Malay language is the major language that is in used by citizen of Malaysia, Singapore and Brunei. As the language is widely used, there are abundant of text or articles in Malay language are available on the internet. This result in the increasing of the articles in Malay language and the number of articles has increased greatly over the years. Thus, the studies for topic selection for Malay articles are very important in order to help clustering the articles into their respective class. In this paper, *k* - Nearest Neighbors (*k* - NN) classifier and Naïve Bayes classifier based approaches were used to classify and assign a topic to the documents according to a predefined topic sets. The approach will be tested by comparing the effects of using different distance method which is the Cosine Similarity and the Euclidean distance on the *k* - NN classifier. Other than that, the effect of stemming on the classifier and the different values of *k* used for the *k* - NN classifier were also tested. In conclusion, the proposed approach had shown that the *k* - NN classifier performs better than Naïve Bayes classifier in performing topic selection for Malay articles. Other than that, the stemming also improves the overall performances of both the classifier in the proposed approach. The findings also show that the application of Cosine Similarity as the distance measure improve the performance of the *k* - NN classifier too.

ABSTRAK

Pemilihan Topik Bagi Artikel Melayu

Bahasa Melayu adalah bahasa utama yang sering digunakan oleh rakyat Malaysia, Singapura dan Brunei. Memandangkan bahasa Melayu telah digunakan secara meluas, terdapat banyak teks atau artikel dalam bahasa Melayu yang boleh didapati di Internet. Hal ini mengakibatkan meningkatnya artikel dalam bahasa Melayu. Oleh itu, kajian untuk pemilihan topik untuk artikel berbahasa Melayu sangat penting untuk membantu pengelompokan artikel ke dalam kelas masing-masing. Dalam laporan ini, pendekatan berasaskan k - Nearest Neighbors (*k* - NN) *dan* Naïve Bayes *digunakan untuk mengklasifikasikan dan menetapkan topik untuk artikel mengikut set topik yang telah ditetapkan. Pendekatan tersebut akan diuji dengan membuat perbandingan kesan ke atas penggunaan pengiraan jarak yang berbeza iaitu antara* Cosine Similarity *dengan* Euclidean *distance ke atas pengklasifikasi k* - NN classifier*. Selain itu, kesan* stemming *kepada pengklasifikasi dan juga nilai k yang diguna untuk k* - NN classifier *juga diuji. Kesimpulannya, pendekatan yang dicadangkan telah menunjukkan bahawa pengelas k* - NN *mempunyai prestasi yang lebih baik berbanding dengan pengelas* Naïve Bayes *dalam pemilihan topik untuk artikel berbahasa Melayu. Selain daripada itu,* stemming *juga dapat meningkatkan prestasi kedua-dua pengelas dalam pendekatan yang dicadangkan. Hasil kajian juga menunjukkan bahawa penggunaan* Cosine Similarity *sebagai pengiraan jarak dapat meningkatkan prestasi pengelas k* - NN.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter discusses and introduces the topic of this report. The problem statement of this project was determined and discussed and the proposed approach, research objective and the proposed methodology were generally described in this chapter.

1.2 Problem Statement

Malay language is one of the languages that are widely used by people in the Southeast Asia especially for those who stay in Malaysia, Singapore, Brunei and also Indonesia and south of Thailand (Sharum, Abdullah, Sulaiman, Murad, Hamzah, 2010). In fact, Malay language is the major language that is in used by citizen of Malaysia, Singapore and Brunei. As the language is widely used, there are abundant of text or articles in Malay language that are available on the internet (Ismail, Saad, Omar, Sembok, 2013). The number of those Malay texts and articles on the internet will continue to increase dramatically as we are in the globalization era. It is important to have topic selection for Malay articles as it is to assist organizing these articles into their respective classes. However, there were very less researches conducted related to the topic selection for Malay articles in the past. Thus, this encourages the studies on topic selection for Malay articles.

Nowadays, the number of documents, articles and text that exist on the internet is so many and in fact it is increasing in a rapid manner. This is due to the fact that Internet has being used to pass information around the world and in most

cases the information were passed around in the form of documents, articles or texts. As the number of those documents and articles is increasing greatly, the efficiency of documents classification, topic selection had become an important task (Ko and Seo, 2000).

There are many researches, models and papers had published to find the best and optimum way to perform topic selection for English articles. However, there were very limited research papers that are focusing on performing topic selection for Malay articles (Alshalabi, Tiun, Omar and Albared, 2013). Thus, this solidifies the reason to further studies on the framework for the topic selection for Malay articles. Alshalabi et al, (2013) had used several machine learning methods with different features selection method to perform the topic selection for Malay articles and had concluded that the *k* - Nearest Neighbor classifier performance better compared to the others classifier. However, the text preprocessing process had being neglected in the experiments done by Alshalabi et al, (2013).

The topic selection is normally done in two steps where first the articles or texts will be pass through the pre-processing where by extraction of the keywords, terms and subjects from the text or articles. The text preprocessing is a vital process for topic selection as it helps to filter out the irrelevant terms or keywords such as the stopwords. The keywords extracted from the text will show the relevancy of the text to the certain keywords. This is because if a particular keyword had made a significant number of appearances in the text, then it is highly possible that the particular keyword represent the topic of the text or articles. Then, the topic selection is a step of selecting from the extracted keywords whether which keywords represent the articles best and suit to be its topic.

Text and article in general has some characteristics such as noise in the data and text structures that are not good. In order to study the text data, the features that represent each word in the text had to be determined (Pramono, Rohman and Hindersah, 2013). The text pre-processing is a step or method where the text will be processed in order to remove the unnecessary terms (Koulali, Mahmoud El-Haji and Meziane, 2013). The typical pre-processing includes the

structural processing, lexical analysis, tokenization, stopwords removal, and stemming and also term categorization (Echeverry-Correa et al., 2014).

1.3 The Proposed Approach

In this paper, the topic selection for Malay articles had being designed by applying the feature extraction which is the text pre-processing phase on the Malay articles and two supervised classification model, *k* - Nearest Neighbor classifier and Naïve Bayes classifier was used to classify the articles into its respective topic. The classification process was done after applying the Term Frequency- Inverse Document Frequency (TF-IDF) as the term weighting metric. It is a metric used to weight the terms obtained after processing the text. The weights, obtained by computing the TF-IDF weight, represent the weighting of the particular word or term in the document sets and it will be used to compute the distance between two documents.

In this paper, the usage of different distance method computation which is the Euclidean distance and the Cosine Similarity as the distance computation were designed in order to investigate its effect on the *k* - NN classifier's performance. The text preprocessing includes the tokenization, punctuation removal and stopwords removal phases. The stemming phases will included in to investigate the impact of the stemming phases in the text preprocessing stage. The topic selection will be done by using the both *k* - Nearest Neighbor (*k* - NN) classifier and Naïve Bayes classifier with the different approach of the text preprocessing and the distance method.

1.4 Research Question

Although it is important to have topic selection to assist in organizing the unstructured text articles, there are limited studies conducted related to topic selection for Malay articles. Many researches in the past were conducted based on English articles by using different approaches of classifier used, text preprocessing task involved, different feature selection method used and many more. This leads

to a main question whether that topic selection for Malay articles is achievable or not. Thus, it leads the 3 following research question that motivate this work.

- Is it feasible to apply topic selection for Malay language articles?
- Is the supervised machine learning techniques effective on topic selection for Malay articles?
- Is the text preprocessing task used for processing English articles applicable for Malay articles?

1.5 Research Objectives

There are three outlined objectives to be achieved in this research paper and the objectives of this research paper are listed below.

- To formulate a standard approach of topic selection for Malay articles and assesses the initial performance of the proposed approach.
- To investigate the effects of using different distance methods on the performance of *k* - nearest neighbor classifier.
- To investigate the effects of applying stemming algorithm as another text preprocessing task to filter the feature extracted from the documents on the classifier.

The proposed framework had being approach by other researchers in the past using other languages document sets. The scope of this research is narrowed down to three main tasks that includes performing the pre-processing task for the Malay document and secondly, to perform a classification task that classifies Malay articles into the appropriate groups before the topic selection process can be performed. Finally, an experimental setup will be designed to assess the performance of the classifier.

1.6 The Proposed Methodology

The following were the methodology involved in this research paper. The dataset is a set of documents collected from various sources which will be described in

Chapter 4. The dataset will be split to two partitions, training and test data. Both training and test data will undergoes the text preprocessing phase where the following task is performed, tokenization, stopwords removal and punctuation removal. Another text preprocessing task which will be performed is the stemming however this task served as one variable where an experiment will be designed to test its effects on the classifier's performances. After the text preprocessing, two classifier, *k* - NN and Naïve Bayes classifier approaches were used to classify the documents.

For the *k* - NN classifier, each document will be given weight by applying the TF-IDF term weighting metric. Then, two distance measure approaches were used to investigate the effect of the distance measure on the *k* - NN classifier's performance. The distance measure that will be used is the Cosine Similarity and the Euclidean distance.

For Naïve Bayes classifier, each of the test documents was computed its probability and the document will be assigned to the class with the highest probability value. Both of the classifier's performances will then be collected, tabulated and analyzed. Both classifiers' performances were measured based on the accuracy of each classifier on each approach. The more detailed methodology will be further discussed in the Chapter 3.

1.7 Organization of the Report

The organization of this report will shows the whole process of the entire research project. A brief and short explanation of the organization of this research is listed below:

Chapter 1: Introduction

This chapter will present the introduction of the research that is being approach which is the topic selection for Malay articles. in this chapter, it includes the further explanation and description regarding the problem statement, objectives of the research project, project scope and the organization of the report.

Chapter 2: Literature Review

In this chapter, the study and review on the other researchers work which are related to the field that being studied in this research paper had been done. The techniques and method that had being used to approach this study by the researchers will be briefly reviewed and studied and will focus more on reviewing the techniques and methods that will be used in this research paper.

Chapter 3: Methodology

This chapter includes the approaches and overall proposed framework used to solve the research question. In this chapter, a brief explanation on how each techniques and methods will be used in designing the proposed framework will be included as well.

Chapter 4: Design Implementation

In this chapter, the implementation of the proposed framework will be implemented by implementing each of the techniques, methods and related algorithms that is mentioned in the Chapter 3.

Chapter 5: Experimental Setup and Results

In this chapter, an experimental setup will be designed to test the performance of the designed framework. The output will be shown and the overall results obtained will be tabulated. The results will be analyzed and discussed at the end of this chapter.

Chapter 6: Conclusion

This chapter will discuss whether the objectives of this paper have been achieved and further discuss the things that can be improved and also the further works that can be performed to improve the quality of this framework in the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In this chapter, the main discussion was focused on the related works done by other researchers in the past. Related works which were related to the topic selection were reviewed and analyzed to understand the trend of the approaches done by the researchers and to have a further understanding of the field which is the topic selection for Malay articles.

2.2 Topic Selection

Topic selection is defined as the task of finding out different themes from the collection of documents. The purpose of topic selection is to find out the topic of the document that exists in the corpus. Topic selection is an important task as it can be used to help to identify, categorize and even labeling the document. It is one of the important tasks in the Information Retrieval field. In a broad sense, topic selection is the task of automatically identifying which of a set of predefined topics are present in the document (Echeverry-Correa et al., 2014). Many approaches have being made to propose an effective and a better framework for the topic selection. One of the examples was the research paper done by Nunes, Kawase, Casanova and Campos (2014) where the authors had approached this study by proposing a combine semantics technologies and statistical method to find, expose and recommend relevant topics as guidance to the forums. In the paper, the proposed method was first performs the Named-Entity Recognition (NER) and topic extraction and then followed by a statistical approach which selects

and ranks the most relevant topic from the forum thread. The diagram below shows the model proposed by the authors (Nunes et al., 2014).

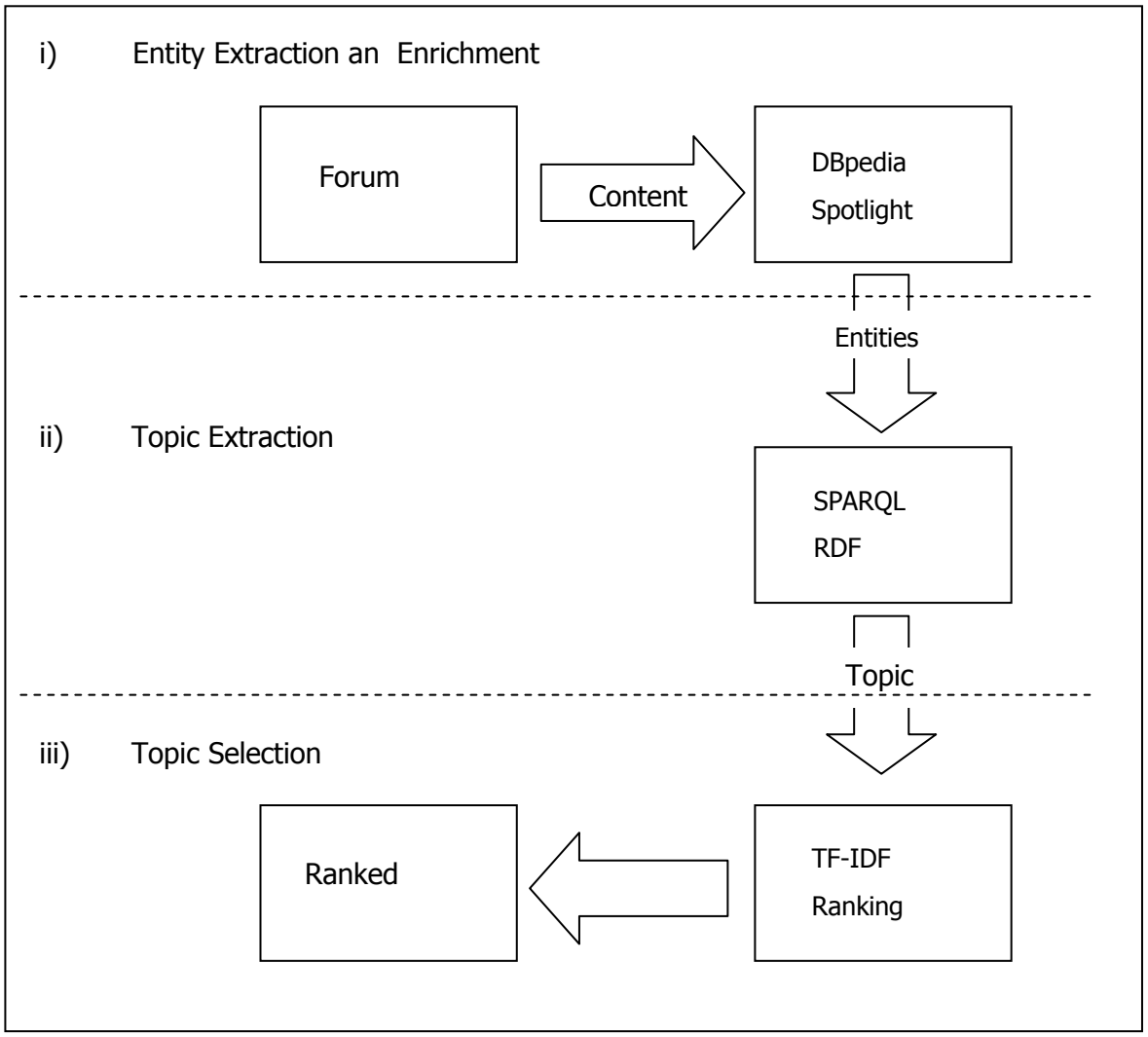

Figure 2.1 Topic Extraction process workflow proposed by Nunes et al. (2014)

Topic selection has become one of the impotant techniques in the studies as it enables the document to be assigned with a topic from the set of predefined topics (Echeverry-Correa et al., 2014). While depending on the research discipline where this task was being carried out, topic selection is also known as topic identification (Echeverry-Correa et al., 2014), topic spotting (Wiener, Pedersen and Weigend, 1995), text categorization (Manning, Raghavan and Schütze, 2008), or

text classification (Yates and Neto, 2011). According to Echeverry-Correa et al. (2014), a conventional topic selection framework consist of the pre-processing, feature extraction, feature selection and classification stages.

However, all the researches mention above were done with English language articles and documents as the test data. There are fewer studies that were done on performing Information Retrieval techniques on Malay Language articles and documents (Alshalabi et al., 2013). Most studies were focused on the text preprocessing techniques, stemming algorithm which will be further discuss later.

Alshalabi et al. (2013) had performed several experiments on the use of feature selection and machine learning methods in Malay text categorization. Alshalabi et al. (2013) had experimented on three machine learning methods which were the *k* - nearest neighbor, Naïve Bayes and N-gram with two feature selection methods which were the Information gain and Chi-square. The results from the experiments performed by Alshalabi et al. (2013) had shown that the *k* - Nearest Neighbor (*k* - NN) classifier achieved the highest performance when compared with the other classifier which were the Naïve Bayes and N-gram classifier. However, the text preprocessing had being neglected in the experiments (Alshalabi et al., 2013).

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 (Khan, Baharudin, Lee, Khan, 2010). The importance of this process can be seen as many similar researches with tasks involving text documents include this process. Samat, Murad, Atan and Abdullah (2008) had included the text preprocessing in the categorization of Malay documents using Latent Semantic Indexing. Samat et al., (2008) mentioned that the pre-processing is basically consists of processes that optimize the list of terms that identify the collection. In the review of machine learning algorithms for documents classification was done by Khan et al., (2010) also had included the feature extraction which is the text preprocessing allow with several

feature selection methods. This show many of the similar researches had also includes the text preprocessing in the process of topic selection as it is a vital process in topic selection as it is to eliminate as much as possible language dependent factors, tokenization, stop words removal and stemming (Khan et al, 2010).

This encourages and motivated the study to show that topic selection can also perform well by applying text pre-processing as the feature extraction on the Malay articles or text documents.

2.3 Text Pre-Processing

Text pre-processing is the first step in a topic extraction/selection process as it reduces the noise in the documents or texts by removing the unnecessary terms (Koulali et al., 2013). Echeverry-Correa et al. (2014) mentioned that the pre-processing stage was important and has substantial impact on the success of the topic identification. The pre-processing steps used were structural processing, lexical analysis, tokenization, stopwords removal, and stemming and also term categorization (Echeverry-Correa et al., 2014). This stage is important as it helps to remove the special characters that exist in the text and removing the stopwords which do not brings meaning to the text and also the stemming of the words to allow the removal of the unnecessary elements or terms in the text. This stage serves as a filter to filter out the noise elements in the text leaving out the terms or keywords that have meaning and weight to the text's content.

However, most of the researches done were using English documents and texts as their document sets. It is questionable that whether can the same approach be used on extracting the keywords from other languages document and text. There are researches done in the past where the document sets used were in languages other than English for example, Arabic, Chinese, Korean, Japanese and many more. An example was the paper done by Bracewell, Yan, Ren and Kuroiwa (2009) where they presented an algorithm for topic analysis which entails category classification and topic discovery and classification. In the experiment, Bracewell et

al. (2009) used the articles in both English and Japanese languages. In the paper conclusion, Bracewell et al. concluded that the presented algorithm works well and able to work with multiple languages and had mentioned that the results obtained from the Japanese articles test were a little worst than the English articles tests where he predicted that it was possibly due to the Japanese articles contains some Chinese characters, kanji thus making the naïve word matching difficult (Bracewell et al., 2009). The following includes the three pre-processing task that will be included in this paper.

2.3.1 Tokenization

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. Echeverry-Correa et al. (2014) mentioned that 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. This process is important as it break down the sentences to tokens where each token is a term or keyword so that further pre-processing task such as stopwords removal and stemming can be done. Example of tokenization is shown in the figure below.

REFERENCE

Argus T. Kwee, Flora S. Tsai and Wenyin Tang, 2009. Sentence-Level Novelty Detection in English and Malay.

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee and Khairullah Khan, 2010. A Review of Machine Learning Algorithms for Text-Documents Classification.

Baeza-Yates, R. A., and Ribeiro-Neto, B. A. 2011. Modern Information Retrieval (2nd edition.). Pearson Education Ltd.

Bernardo Pereira Nunes, Ricardo Kawase, Besnik Fetahu, Marco A. Casanova, and Gilda Helena B. de Campos. 2014. Educational Forums at a glance: Topic extraction and selection.

David B. Bracewell, Jiajun Yan, Fuji Ren and Shingo Kuroiwa. 2009. Category Classification and Topic Discovery of Japanese and English News Articles.

Erik Wiener, Jan O. Pedersen, Andreas S. Weigend, 1995. A Neural Network Approach to Topic Spotting.

Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, and Mohammed Albared, 2013. Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization.

Harun Uguz, 2011. A Two-Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorithm.

J.D. Echeverry-Correa , J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba and J.M. Montero. 2014. Topic Identification Techniques Applied to Dynamic Language Model Adaptation for Automatic Speech Recognition.

Jeong-Ho Chang, Jae Won Lee, Yuseop Kim, and Byoung-Tak Zhang. 2002. Topic Extraction from Text Documents Using Multiple-Cause Networks.

JunChoi Lee, Rosita Mohamad Othman and Nurul Zawiyah Mohamad. 2013. Syllable-based Malay Word Stemmer.

Li Zhixing, Xiong Zhongyang, Zhang Yufang, Liu Chunyong and Li Kuan, 2010. Pattern Recognition Letters.

Liangxiao Jiang and Harry Zhang, 2005. Learning Instance Greedily Cloning Naïve Bayes for Ranking.

Luthfan Hadi Pramono, Arief Syaichu Rohman, and Hilwadi Hindersah. 2013. Modified Weighting Method in TF*IDF Algorithm for Extracting User Topic Based on Email and Social Media in Integrated Digital Assistant.

Man Lan, Chew Lim Tan, Jian Su, Yue Lu, 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization.

Mangalam Sankupellay and Subbu Valliappan. 2006. Malay-Language Stemmer.

Manning, C. D., Raghavan, P., and Schütze, H. 2008. Introduction to Information Retrieval. Cambridge University Press.

Meenakshi and Swati Singla, 2015. Review Paper on Text Categorization Techniques.

Mohd Yunus Sharum, Muhammad Taufik Abdullah, Md Nasir Sulaiman, Masrah Azrifah Azmi Murad, Zaitul Azma Zainon Hamzah, 2010. MALIM – A New Computational Approach of Malay Morphology.

Nordianah Ab Samat, Masrah Azrifah Azmi Murad, Muhammad Taufik Abdullah, Rodziah Atan, 2008. Malay Documents Clustering Algorithm Based on Singular Value Decomposition.

Normaly Kamal Ismail, Nur Hamizah Mat Saad. Sidi Bukhari Sidi Omar, Tengku Mohammad Tengku Sembok, 2013. 2D Visualization of Terms and Documents in Malay Language.

P. Viswanath and T. Hitendra Sarma, 2011. An Improvement to *k* - Nearest Neighbor Classifier.

Qu Chao, Yuan Ruifen, Wei Xiaorui, 2013. *k* - NNC: An Algorithm for *k*-Nearest Neighbor Clique Clustering.

Rim Koulali, Mahmoud El-Hajy and Abdelouafi Meziane. 2013. Arabic Topic Detection using Automatic Text Summarisation.

Roberto H.W. Pinheiro, George D.C. Cavalcanti, Renato F. Correa and Tsang Ing Ren. no year. A Global-Ranking Local Feature Selection Method for Text Categorization.

S. K. Thakur and V. K. Singh, 2014. A Lexicon Pool Augmented Naïve Bayes Classifier for Nepali Text.

Shengyi Jiang, Guansong Pang Meiling Wu and Limin Kuang. 2012. An Improved *k*-Nearest-Neighbor Algorithm for Text Categorization.

Sungjick Lee and Han-joon Kim. 2008. News Keyword Extraction for Topic Tracking.

Syed Abdullah Fadzli, A Khairani Norsalehan, I. Ahmad Syarilla, Hassan Hasni, M Satar Siti Dhalila. 2012. Simple Rules Malay Stemmer.

Tengku Mohd T. Sembok, Zainab Abu Bakar and Fatimah Ahmad. 2011. Experiments in Malay Information Retrieval.

Wei Yong-qing, Liu Pei-yu and Zhu Zhen-fang, 2008. A Feature Selection Method based on Improved TFIDF.

Xiaofei Zhou, Yue Hu and Li Guoa. 2014. Text Categorization Based on Clustering Feature Selection.

Yashodhara Haribhakta, Arti Malgaonkar and Parag Kulkarni. 2012. Unsupervised Topic Detection Model and Its Application in Text Categorization.

Zhengchang Qin, 2006. Naïve Bayes Classification Given Probability Estimation Trees.