Mining Opinion in Online Messages

Norlela Samsudin
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA
Terengganu, Malaysia

Abdul Razak Hamdan
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Mazidah Puteh
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA,
Terengganu, Malaysia

Mohd Zakree Ahmad Nazri
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
Bangi, Malaysia

***Abstract*— The number of messages that can be mined from online entries increases as the number of online application users increases. In Malaysia, online messages are written in mixed languages known as 'Bahasa Rojak'. Therefore, mining opinion using natural language processing activities is difficult. This study introduces a Malay Mixed Text Normalization Approach (MyTNA) and a feature selection technique based on Immune Network System (FS-INS) in the opinion mining process using machine learning approach. The purpose of MyTNA is to normalize noisy texts in online messages. In addition, FS-INS will automatically select relevant features for the opinion mining process. Several experiments involving 1000 positive movies feedback and 1000 negative movies feedback have been conducted. The results show that accuracy values of opinion mining using Naïve Bayes (NB), k-Nearest Neighbor (kNN) and Sequential Minimal Optimization (SMO) increase after the introduction of MyTNA and FS-INS.**

Keywords—Opinion mining; text normalization; feature selection.

I. INTRODUCTION

It was reported on 30th of Jun 2011, 60.7% or 17.7 million Malaysians used Internet. Facebook is the most favored application [1]. Other than that, communication sites such as blogger.com, mudah.com and Twitter were among the top 10 applications that Malaysians used on the Internet [2]. There is a massive amount of information or opinion that can be gathered from these applications. Nevertheless, very few studies had been conducted to mine opinion from messages that are posted online by Malaysians. The following list demonstrates examples of these messages.

- *"oh bestnya, best giler serius. Nak kasik 5 bintang plus2"*
- *Aku bg 4.9 out of 5stars.Yg 0.1 xcukup to sbb aku xfaham.. masa aku tgk aritu pon xfull"*
- *Ksian ngan kawan aku. Coz abihkan duit utk film nih"*

The examples indicate the following characteristics:

- The use of Malay and English words with Malay words as the main contributors. This scenario is known as Bahasa *Rojak*.
- There is a high number of abbreviations such as *sbb*, *bg* and *tgk*.
- The sentences do not follow the correct syntax of sentence development.

The above scenario make it difficult to mine opinion using natural language processing as expressed by [3]

> "*One drawback of an NLP based approach is that it would likely perform very poorly when used on grammatically incorrect text… methods to detect and possibly correct bad English would be necessary before use on a large scale*."

Furthermore, recognizing subjective words that are relevant to opinion is also a problem in mining opinion using the machine learning approach. The current feature selection techniques in machine learning approach such as Document Frequency (DF), Chi Square and Information Gain assign a value to each feature based on a particular statistical equation.

The features are then sorted. It is up to the user to select the appropriate features based on the sorted value. Different users may select different features. Often, a newbie who is not aware of this scenario would do nothing and causes the classifier transaction to take a longer processing time and use more resources.

The objective of this paper is to introduce a method to normalize noisy texts in Mixed Malay Language texts with the introduction of Malay Mixed Text Normalization Approach (MyTNA). In addition, a new feature selection method named Feature Selection based on Immune Network System (FS-INS) is introduced to select relevant features in opinion mining.

The remainder of the paper is organized as follows: In Section II, previous works in opinion mining using machine learning approach, normalization of noisy text and feature selections in opinion mining process are reviewed. The MyTNA steps and FS-INS algorithm are clarified in Section III. The performance of FS-INS is discussed in Chapter IV. Lastly, conclusion of the study and future research direction are explained in Section V.

II. BACKGROUND

A. Opinion Mining using Machine Learning Approach

Opinions, beliefs, emotions and sentiments are part of private states that cannot be observed. These states are expressed in a document using subjective words [4]. Subjective words that identify the private states may be identified using specific dictionary such as WordNet or SentiWordNet. At the beginning of this century, Pang, Lee and Vaithyanathan [5] started using machine learning approach to mine opinion. Prior to that, opinion mining activities were carried out using natural language processing (NLP) approaches ([6] [7] [8]). Pang, Lee and Vaithyanathan [5] successfully used text mining activities in mining opinion from 700 positive and 700 negative movie reviews. They concluded that additional activities to identify sentiment were required in opinion mining using the machine learning approach. Several researchers used NLP activities in pre-processing steps to select features that are relevant to sentiments ([9] [10] [11]). Other than that, Pang and Lee [12] utilized statistical technique to identify a sentiment phrase. Sentences without sentiment phrases were removed before the opinion mining process. Similarly, Barbosa and Feng [13] used items such as icons and the existence of sentiment words to identify sentiment phrases. Clearly, additional activities in addition to the normal text mining processes are required in opinion mining process. Even though the number of research activities on opinion mining has increased for the past century, none of them studies the performance of opinion mining in Malay language or in bahasa rojak. That is the objective of this paper.

B. Previous works in normalization of noisy texts

Knoblock, Lopresti, Roy and Subramaniam [14] define noisy texts as "any kind of difference between the surface form of a coded representation of the text and the intended, correct or original text". Before the year 2000, most works on normalization of noisy text involved documents that were created using OMR ([15] [16]). Normalization of noisy texts in short message service (SMS) started to appear in 2005 ([17] [18]). Lately, the normalization of noisy texts has started to use data from online applications such as Twitter messages and Facebook entries. ([19] [20] [21]).

In general, there are three ways to execute the normalization process i.e correction of spelling, machine translation and automatic identification of phonetic. This study uses the first method which includes identifying a noisy text, finding the candidate of correct terms and selecting the correct term.

C. Previous works in feature selection of opinion mining

Feature selection is the process of selecting a set of attributes or features that is relevant to the mining processes. In relation to text mining or opinion mining, every distinct word that exists in the corpus is considered as a feature. The traditional method of feature selection is by selecting all features in a method known as bag of words (BOW). Unfortunately, this method causes certain classifier to perform poorly due to high requirement of resources and longer execution time. Therefore selecting relevant features without reducing the performance of opinion mining process is important. Previous researches in opinion mining use two approaches of feature selection. The first approach uses NLP processes such as Part of Speech (POS) in identifying certain sentence structure or stemming and lemmatization transaction to reduce related forms of a word to a common base form ([5] [9] [22] [23]). The second approach assigns a specific value to every features based on certain statistical equation. Document Frequency uses frequency of the words that exist in the corpus. Information Gain and Mutual Ratio use probability of a word occurring in each class and Chi Square calculates the degree a word is not relevant to a particular class. Unfortunately, these statistical techniques assign a value and sort the features based on these values. It is up to the users to indicate which features should be selected. This study introduces a new feature selection technique that will calculate and select the feature automatically.

III. METHODOLOGY

Fig. 1 illustrates the opinion mining process with the introduction of MyTNA and FS-INS. Both the training data and test data went through normalization process before the opinion mining process.

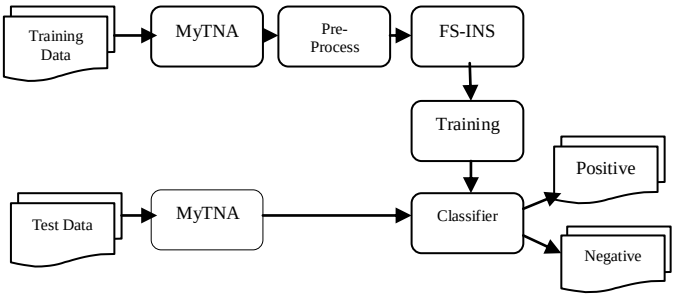

Fig. 1. Opinion mining process for Malay Mixed language

A. Malay Mixed Text Normalization Approach (MyTNA)

The objective of MyTNA is to reduce the number of features by correcting the spellings of common noisy terms and abbreviations that exist in online messages. For example, the word 'tidak' is written as 'tak', 'x', 'dok', and 'dak' in online messages. When these words are transformed to the features in opinion mining process, there will be five different features that represent the word 'tidak'. These scenario influences calculation of several classifier such as Naïve Bayes that uses probability calculation in the classification process. Other than that, the value of the word 'tidak' in the calculation of k-Nearest Neighbour classifier will be very low since the frequency of the words is 1 instead of 5. Additionally, the word 'tidak' is considered as irrelevant if the frequency is low and divided into five different terms instead of one. Therefore, incorrect spellings of words in online messages have to be corrected before the opinion mining process is executed. In this study, the correction of spellings was done using MyTNA

A corpus that consists of 21,000 randomly extracted online messages was derived from e-forum, Facebook and Twitter entries. The following lists were constructed based on this corpus:

This research is supported in part by the Fundamental Research Grant Scheme (FRGS) under the ninth Malaysia Plan (RMK-9), Ministry of Higher Education (MOHE) Malaysia. The grant no is 600-RMI/ST/FRGS 5/3/Fst (208/2010).

- A list of common noisy terms, which consists of noisy terms that exists more than 5 times in the corpus.
- A Bi-Gram list, which consists of the frequency of a word that exists with another word in the corpus.
- A list consists of common English words that are used in the online messages written by the Malaysian. Digital English dictionary is not used in this study because of high similarity between noisy terms in Malay language and the English word such as *'cite'* and *'die'*.

Other than these lists, a list of artificial abbreviations was also derived using the rules that were explained in [24]. During normalization process, each word in the message was tested for out of vocabulary word using a digital Malay dictionary, Common English Word list and Common Acronym list. If the word did not exist in any of these lists, it is considered as a noisy term. The Common Noisy Term list and Artificial Abbreviation list were used to identify potential candidates of the correct word. Later, the Bi-gram list was used to determine the correct translation of the noisy term. MyTNA steps were summarized in Fig. 2.

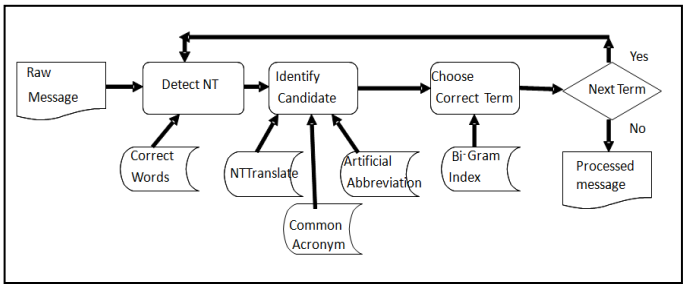

Fig. 2. MyTNA steps

B. Pre-processing

In pre-processing step, features that were not relevant to sentiment were eliminated. The following activities were executed in this study.

- All uppercase letters were changed to small case letters. This is because, online users tend to be creative in writing a message and might write the word 'best' as 'Best', 'BEST', 'BeSt , 'BeST', and 'BesT'. Changing all the uppercase letters into small case letters reduce the number of features.
- All stop words (words without meaning) for Malay language and English language were removed.
- The word 'tidak' was used with the subsequent word. Normally, the word 'tidak' indicates a sentiment. Phrases such as 'tidak best' and 'tidak suka' are synonym with negative sentiments. The word 'tidak' itself exists in positive and negative documents. Normally, the frequency of the word 'tidak' is high in both classes and may be irrelevant to the class. Therefore, using the word 'tidak' with the subsequent word will reduce the frequency of the word 'tidak' itself and increase the number of words that represents the sentiments.

C. Feature Selection based on Immune Network System (FS-INS)

In the filter typed feature selection approach, related features were selected based on certain mathematical equation. Each feature was given a value and later sorted accordingly. However, related features were not selected automatically. Therefore, the users would have to check which features to be selected. If the users were not aware of the activity, all features would be selected, hence resulting poor execution time or the system to crash due to insufficient resources. In term of opinion mining, selecting features that are relevant to positive and negative sentiments is also important. Therefore, in this study, each feature was given a value based on a formula that was introduced by Simeon and Hilderman [25] named Categorical Proportional Difference (CPD). Keefe [26] adjusts the formula to fit the two classes case as shown in Equation 1.

(1)

$$CPD(t) = \frac{|FP_i - FN_i|}{FP_i + FN_i}$$

Where

- FP_i is the frequency of term (t) exists in the positive class;
- FN_i is the frequency of term (t) exists in the negative class;

This formula considers a feature as relevant if it occurs in one class in a higher frequency than its frequency in any other classes. For example the feature 'bagus' that exists 100 times in the positive class and 10 times in the negative class will be assigned a value 0.82. On the other hand, a feature that exists in the same frequency in the positive class and negative class is regarded as irrelevant. For example, the term 'saya' that exists 100 times in both classes will be assigned CPD value as 0.0. Therefore, a high CPD value means the feature is very relevant and should be selected. Unfortunately, if a feature exists in only one class, the value 1.0 is assigned to the feature regardless its frequency. Therefore, the relevancy of a feature that exists only once is considered the same as a feature that exists 100 times in only one classAnother problem is to identify the limit of features that is considered as relevance. . In the traditional feature selection techniques, this value is identified by the user based on his or her interpretation of relevancy. To solve these problems, a new feature of selection algorithms based on artificial immune network named FS-INS was created. The main characteristics of this algorithm are listed in the following list:

- If a feature exists in only one class, only the feature that exists more than a certain threshold would be selected.
- A feature is considered as relevant if its CPD's value is above certain threshold.
- A feature is considered relevant if its CPD's value is similar to other features. A feature with CPD's value that matches many other features with similar CPD's value has higher relevancy as compared to other features with less matched of CPD's value.

In the artificial immune system, a feature is considered as a B cell. A memory is used to choose cells with similar CPD's value, while a cell with less matched CPD's value will be deleted from the memory. At the end of the process, all B cells in the memory are considered as the selected features.

D. Experiment's Data

Data for the experiments were randomly collected from various online forums used by Malaysian users such as http://mforum.cari.com.my/portal.php and http://www.mesra.net/forum/. The data were also retrieved from Facebook entries and Twitter messages. Online messages that were selected have the following characteristics.

- It has feedback on a particular movie;
- It has a positive or a negative sentiment; and
- It is written in Malay Mixed Language.

1000 positive movie feedbacks and 1000 negative movie feedbacks were collected and used in all experiments.

E. Experiments

Opinion mining process was executed to analyze the efficiency of MyTNA and FS-INS. k-Nearest Neighbour was used as the classifier in these experiments. Several values of k were tested and it was found that the accuracy value was at the highest in most cases when it is set to 1. Therefore to ensure its consistency, k was set to 1 in all of the experiments. Other than kNN, the accuracy value using Naïve Bayes and Sequential Minimal Optimization were also collected. The accuracy of each opinion mining process was calculated. The accuracy value was calculated using formula in Equation 2

$$Accuracy = \frac{\#\ of\ correct\ prediction}{\#\ of\ reviews} \quad (2)$$

The following experiments were conducted using Weka 3.6 application as the opinion mining tool. Table 1 summarized the experiments for easy understanding.

E1: Opinion mining using raw data;
E2: Opinion mining using raw data and pre-process activities;
E3: Opinion mining using raw data and FS-INS;
E4: Opinion mining using data which had been normalized using MyTNA activities (processed data);
E5: Opinion mining using processed data and pre-process activities;
E6: Opinion mining using processed data, pre-process activities and FS-INS.

TABLE I. LIST OF EXPERIMENTS

Eksp.	Raw Data	MyTNA	Pre Process	FSINS
E1	√			
E2	√		√	
E3	√			√
E4		√		
E5		√	√	
E6		√	√	√

IV. ANALYSIS OF RESULT

The objective of this study is to improve the result of opinion mining messages that are written in 'Bahasa Rojak'. Therefore, the normalisation of noisy text through MyTNA activities and reduction of features using FSINS were applied in the opinion mining process. In addition, several pre-process activities were also conducted prior to the feature selection step. Several experiments were conducted to check the efficiency of these new steps. Table 2 illustrates the result of these experiments.

TABLE II. RESULTS OF EXPERIMENTS

Eksp.	Accuracy		
	NB	*kNN*	*SMO*
E1	85.00	64.10	82.96
E2	87.20	68.00	85.60
E3	86.90	66.70	82.63
E4	85.80	64.07	81.96
E5	89.60	72.60	86.80
E6	91.04	79.08	92.25

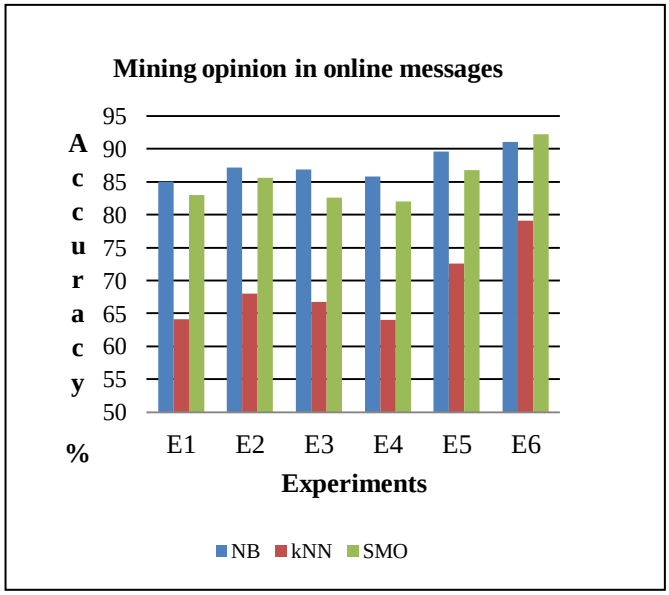

Fig. 3. Results of Experiments

The following conclusions were derived from the results.

- Normalisation of noisy text alone does not improve the opinion mining result. (The result of Experiment E4 is similar to the result of Experiment E1).
- Using only FSINS also does not improve the opinion mining result (The result of Experiment E3 is similar to the result of Experiment E1).
- Combining normalisation of noisy text and the pre-processing activities improves the result of opinion mining slightly (The result of Experiment E5 is better than the result of Experiment E1).
- Combining normalisation of noisy text, the pre-processing activities and using FS-INS as feature selection improves the accuracy value of opinion mining in mixed Malay language (5 % in NB, 15% in kNN and 9% in SMO as shown in Fig. 3.)

It can be concluded that choosing the relevant features improves the result if opinion mining and NB used the probability calculation to predict a class. Additionally, selecting relevant features also improves the probability for predicting the class of new online messages. Similarly, k-Nearest Neighbour classifier uses the class of the nearest neighbour in its prediction. The normalisation of noisy texts process and FS-INS corrects the spelling of most words. In addition, the reduction of features in pre-processing steps and the feature selection technique make it easier for k-Nearest Neighbour to predict the class of a particular message. SMO classifier also creates a virtual line between both classes. Relevant features cause a better line prediction and lead to better accuracy in predicting the appropriate class.

Pang, Lee and Vaithyanathan [5] indicates that additional activities to the normal activities of text mining are required in opinion mining. In this study, several activities were introduced to ensure that only relevant features to sentiments are selected such as

- using of word ‘tidak’ with the subsequent word;
- selection of features based on CPD’s values; and
- selection of features with similar CPD’s values.

These activities contribute to the improvement of accuracy value in opinion mining of online messages written in Malay Mixed Language.

V. CONCLUSION

Improving the accuracy of opinion mining of online messages written in ‘Bahasa Rojak’ is the objective of this study. Executing additional activities such as normalisation of noisy texts approach named MyTNA, several pre-processing activities and a feature selection technique named FS-INS improve the result of opinion mining using NB, kNN and SMO as the classifiers. Nevertheless, more experiments are required to verify whether additional activities introduced in this study improve the opinion mining process. One of them is to validate the result of using FS-INS as feature selection technique as compared to the result of opinion mining using other feature selection techniques such as Document Frequency, Information Gain and Chi Square.

REFERENCES

[1] http://www.internetworldstats.com/stats3.htm, accessed 1/4/2013

[2] http://www.alexa.com/topsites/countries;0/MY, accessed 1/4/2013

[3] O'Neill, A.: ‘Sentiment Mining for Natural Language Documents’, Book Sentiment Mining for Natural Language Documents’, Australian National University, 2009

[4] Wilson, T.W., J.: ‘Annotating Opinions in the World Press’, 2003

[5] Pang, B., Lee, L., and Vaithyanathan, S.: ‘Thumbs up?: sentiment classification using machine learning techniques’. In Proc. of the ACL-02 conference on Empirical methods in natural language processing, 2002, pp. 79-86

[6] Turney, P.D.: ‘Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews’. In Proc of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002 pp. 417-428

[7] Nasukawa, T., and Yi, J. ‘Sentiment analysis: capturing favorability using natural language processing’. in Proc. of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 2003 pp. 70-77

[8] http://nlp.stanford.edu/courses/cs224n/2007/fp/johnnyw-hengren.pdf, accessed 21 January 2010 2010

[9] Gamon, M., ‘Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis’. In Proc. of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004 , pp. 841

[10] Boiy, E., and Moens, P.H. ‘Automatic Sentiment Analysis in On-line Text’, ‘Book Automatic Sentiment Analysis in On-line Text’ 2007, pp. 349-360

[11] Boiy, E., and Moens, M.-F.: ‘A machine learning approach to sentiment analysis in multilingual Web texts’, Information Retrieval, 2009, 12, (5), pp. 526-558

[12] Pang, B., and Lee, L., ‘Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales’. In Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, 2005 pp.115-124

[13] Barbosa, L., and Feng, J.: ‘Robust sentiment detection on Twitter from biased and noisy data’. In Proc. of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China 2010 pp. 36-44

[14] Knoblock, C., Lopresti, D., Roy, S., and Subramaniam, L.: ‘Special issue on noisy text analytics’, International Journal on Document Analysis and Recognition, 2007, 10, (3), pp. 127-128

[15] Kernighan, M.D., Church, K.W., and Gale, W.A.: ‘A spelling correction program based on a noisy channel model’. In Proc. of the 13th conference on Computational linguistics - Volume 2, Helsinki, Finland, 1990, pp. 205-210

[16] Mikheev, A.: ‘Document centered approach to text normalization’. In Proc. of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece, 2000 pp. 136-143

[17] Aw, A., Zhang, M., Xiao, J., and Su, J., ‘A phrase-based statistical model for SMS text normalization’. In Proc. of the COLING/ACL, Sydney, Australia, 2006 pp. 33-40

[18] Fairon, C.P., S.: ‘A Translated Corpus of 30,000 French SMS’, 2005

[19] Clark, E., and Araki, K.: ‘Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English’, Procedia - Social and Behavioral Sciences, 2011, 27, (0), pp. 2-11

[20] Contractor, D., Kothari, G., Faruquie, T.A., Subramaniam, L.V., and Negi, S.: ‘Handling noisy queries in cross language FAQ retrieval’. In Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, 2010 pp. 87-96

[21] Laboreiro, G., Lu, Sarmento, s., Teixeira, J., Eug, and Oliveira, n.: ‘Tokenizing micro-blogging messages using a text classification approach’. In Proc. of the fourth workshop on Analytics for noisy unstructured text data, Toronto, Canada, 2010 pp. 81-88

[22] Dave, K., Lawrence, S., and Pennock, D.M.: 'Mining the peanut gallery: opinion extraction and semantic classification of product reviews'. In Proc. of the 12th international conference on World Wide Web, Budapest, Hungary, 2003 pp. 519-528

[23] Salvetti, F., Reichenbach, C., and Lewis, S.: 'Opinion Polarity Identification of Movie Reviews Computing Attitude and Affect', in Text: Theory and Applications', in Shanahan, J., Qu, Y., and Wiebe, J. (Eds.) (Springer Netherlands, 2006), pp. 303-316

[24] [Samsudin, N., Puteh, M., Hamdan, A.R., and Ahmad, M.Z.: 'Normalization of Common NoisyTerms in Malaysian Online Media','In Knowledge Management International Conference (KMICe) , Johor Baharu, 2012, pp. 526 - 531

[25] Simeon, M., and Hilderman, R. 'Categorical Proportional Difference: A Feature Selection Method for Text Categorization', in Proc. of Conferences in Research and Practice in Information Technology (CRPIT), 2008, pp. 201-208.

[26] Keefe, T.O.K., I. 'Feature Selection and Weighting Methods in Sentiment Analysis', in Proceedings of the 14th Australasian Document Computing Symposium, 2009, pp. 67-74.