Immune Based Feature Selection for Opinion Mining

Norlela Samsudin, Mazidah Puteh, Abdul Razak Hamdan, Mohd Zakree Ahmad Nazri

***Abstract*— Opinions about a particular product, service or person are communicated effectively through online media such as Facebook, MySpace and Twitter. Unfortunately only a few researchers had researched on the performance of opinion mining using online messages that were written in Malay Languages. Opinion mining processing that uses Natural Language Processing approach is difficult due to the high content of noisy texts in online messages. On the other hand, opinion mining that uses machine learning approach requires a good feature selection technique since the current filter typed feature selection techniques require interference from the user to select the appropriate features. This study used a feature selection technique based on artificial immune system to select the appropriated features for opinion mining. Experiments with 2000 online movie reviews illustrated that the technique has reduced 90% of the features and improved opinion mining accuracy up to 15% with k Nearest Neighbor classifier and up to 6% with Naïve Baiyes classifier.**

***Index Terms*—Data Mining, Opinion Mining, Feature Selection, Artificial Immune System**

I. Introduction

Opinion mining is a process of gathering sentiments or opinions from a group of documents. Nowadays, it is very easy to express or gather opinions about a particular product, service or people. Online applications such as online forums, SMS, Twitter and Facebook are widely used as medium of communication to convey these messages. Therefore, it is important for an organization to gather the sentiment of their customers when they consumed a product or a service through the opinion mining activities. In fact the advancement of technology and online communication media has encouraged more researches on opinion mining activities. Nevertheless, very few works had been carried out involving opinions which were written online by the Malaysians.

The high contents of noisy texts in the online messages contributed to the low number of researchers. In addition, the messages normally did not follow the correct rules and grammar that had been established in creation of sentences. Other than that, online messages created by the Malaysians incorporated words from many languages such as words from the Malay language, words from the English language and words that have been used locally which are known as local dialect/slang. These problems cause Natural Language Processing (NLP) activities such as Part of Speech (POS) and lemmatization impractical. On top of that, until recently, there is no dictionary which is similar to WordNet or SentiWordNet that is written in the Malay Language. Therefore identifying sentiment words in Malay words is difficult without these references.

Another approach in opinion mining is by using machine learning technique, which utilizes text classification activities. Using machine learning approach has another effect. A good feature selection technique is required to reduce the number of features in the opinion mining process.

This paper explains a feature selection technique named Feature Selection based on Immune Network System (FS-INS). Inspired from the Artificial Immune System (AIS) theory, experiments showed that the technique was better than the traditional 'filter' typed feature selection techniques i.e. Document Frequency (DF), CHI Square (CHI) , Categorical Proportional Difference (CPD) and Information Gain (IG).

The remainder of the paper is organized as follows: In Section II, the concept of AIS is summarized. In addition, a few previous works on applying feature selection on opinion mining processes are also reviewed. The FS-INS algorithm and the dataset that were used in the experiment are clarified in Section III. The performance of FS-INS in comparison to the performance and other feature selection techniques is reviewed in Chapter IV. Lastly, conclusion of the experiment and future research direction are explained in Section V

II. Background

A. Feature Selection with Opinion Mining

Feature selection is an activity, which select relevant features based on a particular measurement. Its main purposes are to simplify the training process and to reduce the time of training process. The performance of certain classifier such as k Nearest Neighbor is poor when the features are too many. On the other hand, it is important to select a feature selection technique which reduces the number of features without reducing the performance of opinion mining. In previous researches on opinion mining, several common feature selection techniques such as POS [1] [2], Information Gain [3], Document Frequency [4] and

Manuscript received Mach 18, 2013; revised April 10, 2013. This work was 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).

Norlela Samsudin is with Faculty of Computer and Mathematical Science, Universiti Teknologi MARA Terengganu, Malaysia, norlela@tganu.uitm.edu.my

Mazidah Puteh is with Faculty of Computer and Mathematical Science, Universiti Teknologi MARA Terengganu, Malaysia mazidahputeh@tganu.uitm.edu.my

Abdul Razak Hamdan is with Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
arh@ftsm.ukm.my

Mohd Zakree Ahmad Nazri with Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
mzan@ftsm.ukm.my

Chi Square [5] were incorporated.

There are three groups of feature selection techniques i.e. filter, wrapper and embedded. In a filter category, a group of features is selected based on a particular mathematical equation and may be used with any classifier. Contrary to that, the features that are selected in the wrapper and the embedded techniques are bound to a particular classifier. Other than very rigid in term of classifier, the wrapper and embedded techniques normally require high allocation of resource and require longer execution time.

DF, CHI and IG are examples of feature selection techniques that falls under the filter category. DF counts the frequency of a particular word in all categories. Features on the top and bottom of the list are then removed. CHI calculates the degree where a particular feature is not relevant to a particular class. Features are sorted based of its relevance to a particular class. IG calculates the relevance of a feature based on the probability of the word exist in a particular class. Another feature selection method that was included in this study is Categorical Proportional Difference (CPD). It is a measurement that takes into consideration the existence of a feature in a class in its calculation. It is introduced by [6] and used by several other researchers such as [7] and [8] . The calculation which is used in [7]

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

was adopted in this study. The formula is written in (1).

(1)

Where:
FP_i is the number of feature i in the positive class
FN_i is the number of feature i in the negative class
DF_i is the number of document where feature i exist.

A feature is considered to be relevant to a particular class if its existence is high in a particular class in comparison to its existence in another class. For example, the value of CPD (*bagus*) that exists in 10 positive documents and 2 negative documents is 0.67. Similarly the value of CPD *(tidakbagus*) that exists once in a positive document and exists in 8 negative documents is 0.78. On the other hand, the value of CPD (*saya)* which exists in 5 positive documents and 4 negative documents is 0.11. The lower value shows that feature '*saya*' is not relevant to either the positive or the negative class or not relevant to the opinion mining process.

Even though the filter typed feature selection is simple and flexible, it requires human interaction since the current technique will sort the features based on specific relevance measurement. It is up to the user to select a group of features based on the user's observation of the calculated values which are different from one set of data to another set. If the user fails to set the limit, all features will be used in the classification activities.

B. Artificial immune system

The word immunity refers to the reaction of an organism to infectious diseases by foreign substances known as pathogens or antigens. The immune system consists of cells and molecules that interact with each other, initiate immunological responses and destroy the foreign substances. Lymphocytes which consist of B cells and T cells are the main constitutes of this system. Inspired by the human immune system, AIS has been used to solve problems in many areas such as machine learning, robotic, computer security, clustering, classification and web mining [9]. Castro and Timmis [10] define AIS as

> "adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving."

More comprehensive reviews on the biological aspects and scientific aspects of AIS were written in several literatures such as [11] and [12].

Since the beginning of this century, AIS has been used to solve classification and clustering problem. Timmis (2001) utilized AIS technique to solve clustering problem. He proposed that the AIS was able to 'learn' from history data and recognized a specific pattern. Since then, several researchers have employed this technique to solve text classification problems. Twycross (2002) used AIS to classify document into several topics. [13] used AIS to identify uninteresting emails. Spam detection is another area where AIS may be applied ([14], [15]). Other than that, AIS was also used to search for interesting web pages ([16]) and to classify medical documents ([17],[18]).

The following characteristics inspired the used of AIS in this study.

1. **Pattern Recognition** refers to the ability to recognize pattern and maintain the pattern until a new and more dominant pattern emerges. This is possible through the interaction between B cell and the antigen. In addition to that, the use of memory population allows the pattern to be 'remembered'.

2. **Affinity Maturation** refers to how a cell is selected to be cloned and how the mutation process is done. It is important since it will identify the pattern for each class. In this study, a small difference between the CPD values shows similarity between two cells.

3. **Meta dynamics** refers to activities which will control the number of population. A stimulation counter is maintained for each cell in memory population. A B-cell with high number of similarities with other B-cell is kept, whereas B-cell with low stimulation will be deleted from the population. These activities refer to the Negative Selection of AIS theory.

This study did not use all characteristic of AIS. Cloning and mutation concepts were not utilized in this study since they might lead to incorrect evaluation. Changing the spelling of the feature will introduced more features and noisy texts to the population. These noisy texts were cleaned before the mining processes. In addition to that, the results of opinion mining using this technique were better without mutation and cloning module.

III. Methodology

Data Preparation

1000 positive movie feedbacks and 1000 negative movie feedbacks were collected randomly from online forums, Facebook messages and Twitter messages which were created by the Malaysians. Several preliminary works had been executed prior to the opinion mining process. The online data was normalized using techniques explained in [19]. In addition to that, words with no meaning were removed. The stop words list was created from the following sources:

- The English stop words list taken from Armand Dbrahaj's website [20]. The list was selected since the number of words was moderate and the researcher had used it successfully in his research.
- The Malay stop words list was adopted from a list used by [21]. Nevertheless, a few terms that were related to opinion such as '*tidak*' were excluded from the list. In total, there were 540 stop words in the list.

The messages were then converted into the lower case format. These activities were carried out in order to remove redundant words thus reducing the number of features. Lastly, the word '*tidak*' was merged with the next word. Word '*tidak*', which mean '*no*' in English plays a major role in expressing negative sentiment. The word '*tidak*' by itself will carry no meaning since it is one of the words that exists frequently in the positive and negative classes. Combining the word with the next word normally introduces new features that represent sentiment such as '*tidakbagus*' or '*tidakcantik*'. This is a technique to cater for negative words in the Malay language.

Data Representation

The first activity that is recommended by [12] in applying AIS in solving a problem is mapping the concept between the Immune System and the project. Table 1 shows the mapping between Immune System and FS-INS.

Table 1: Mapping of concept between Immune System and FS-INS

Immune System	This Project
Antibody	Features in the population
Antigen	Another feature
Memory of antibody	A group of features which had undergone the training phase.

The next activity as shown in Table 2 is cell representation of the problem solving.

Table 2 : Cell Representation

Element	Explanation
text	Features
FP	The number of the word exist in positive class
FN	The number of the word exist in negative class
sti	The number of stimulation
match	The number of match with other cells
CPD	CPD value

FS-INS Algorithm

Fig. 1 shows the algorithm of FS-INS. It starts with the creation of AG population. In this module data were converted to cells representation and the value of CPD for each cell was calculated. Next, the value of affinity threshold was calculated. The parameter held the average value of affinities among all cells in the population. A number of cells was then selected randomly to initiate the memory population.

```
Start
   Create $AG_{test}$
   Calculate AT
   Initial MC
   for $ag_i$ Ɛ AG
      lowAff ← lowest ( $ag_i$, $mc_i$ Ɛ MC)
      if (lowAff < AT ) dan ($ag_{lowAff}$.CPD > CT)
           bc ← $ag_{lowAff}$
           bc.setSti (ST)
           MC ←MC U bc
           for all $mc_i$  Ɛ MC
              if $mc_i$ . CPD == bc.CPD
                 $mc_i$ . match++;
           endFor
        endif
        for all $mc_i$  Ɛ MC
         if $mc_i$ . sti <= 0 &&  $mc_i$ . match < MT
              MC ← MC - $mc_i$
         endif
         $mc_i$ . sti ← $mc_i$ . sti - 1
       endFor
     endFor
   Classification using $ag_i$ Ɛ MC as features
   Evaluation
 End
```

Fig. 1: The FS-INS Algorithm

Later, for each data in AG population, it will be compared to the cells in memory population. Similar cell with low affinity values were kept in the population. Less important cells with low match values and low stimulation values would be deleted from the memory population. At the end of the process, only relevant cells that remained in the memory population were considered as relevant features and were used in the classification process.

Evaluation

The performance of FS-INS was evaluated by executing the opinion mining process using modules prepared in Weka 3.6 application. The accuracy of opinion mining using FS-INS as feature selection was compared to the accuracy of opinion mining without any feature selection. In addition, the performance of opinion mining of FS-INS is compared to the result of opinion mining with other common feature selection techniques i.e. DF, CHI, IG, and CPD. The accuracy value was calculated using equation in (2).

$$Accuracy = \frac{\#\ of\ correct\ prediction}{\#\ of\ reviews} \tag{2}$$

IV. Results and discussion

Fig. 2 illustrates the reduction of features in experiments which varied the number of data from 100 positive movie reviews and 100 negatives movie reviews until 1000 positive movie reviews and 1000 negatives movie reviews. The result illustrated that the number of features was reduced up to 90% when FS-INS was utilized as the feature selection technique.

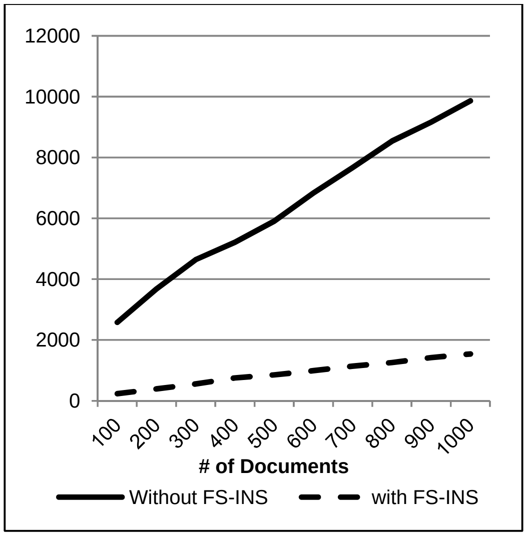

Fig. 2: Reduction of Features

Table 3 illustrates the performance of FS-INS as feature selection in opinion mining with three common classifiers i.e. Naïve Bayes, k Nearest Neighbor (k=1) and Support Vector Machine (SVM).

Table 3: Performance of FS-INS in NB, kNN and SMV Classifier

Classifier	% Acc based	% Acc with FS-INS	% different
NB	84.97	91.04	6.07
kNN	64.1	79.08	14.98
SVM	82.9	92.25	9.35

The result illustrated an increase trend in the performance of opinion mining for all classifiers when FS-INS was used as the feature selection technique. The highest increment was recorded when kNN model was used as classifier. This is expected since kNN works best with small numbers of relevant features. In addition to that, FS-INS chooses cell with similar values which is the same to the nearest neighbor concept in kNN. Similar to the result of opinion mining obtained by [1] and [7] the result of opinion mining with SMV was the highest in comparison to the result of opinion mining with NB and kNN.

Fig. 3 illustrated the results of opinion mining using NB as the classifiers. The results of opinion mining using FS-INS as feature selection were compared to the results of opinion mining using other feature selection techniques such as DF, CHI, IF and CPD. The performance of CPD was expected to be very poor.

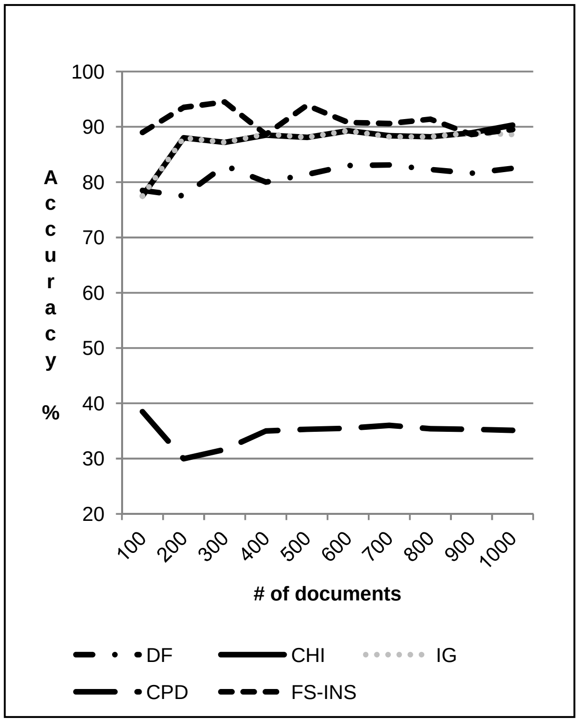

Fig. 3: Result of OM using NB as Classifier

Similar to the other filter typed feature selection techniques, all features were sorted based on the CPD values. When a feature existed in only a class and did not exist in another class a value of 1 was assigned. The value assigned by CPD to a feature that existed 20 times in a particular class and none in another class was very relevant and was the same to a feature that existed once in a particular class and none in other classes.

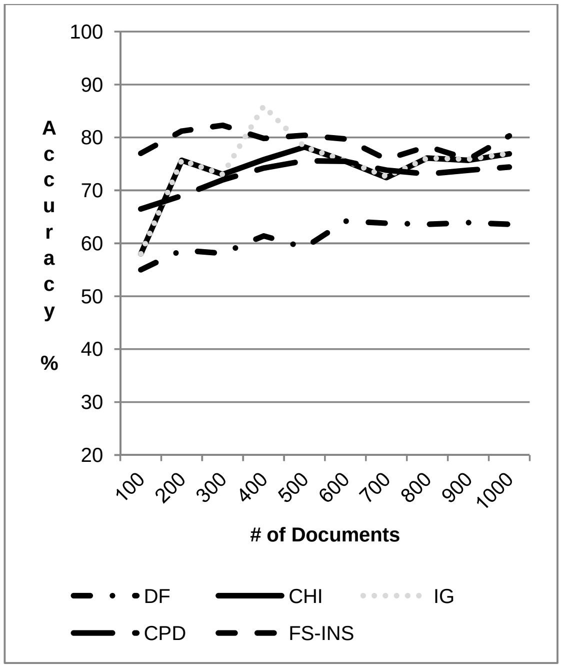

Fig. 4 : The performance of OM with k-NN as the classifier

Therefore, there were chances that features which were not relevant to the class were selected. NB classifier used probability measurement to predict the class of a new document. When irrelevant features were selected in feature selection activities, the accuracy of opinion mining dropped tremendously.

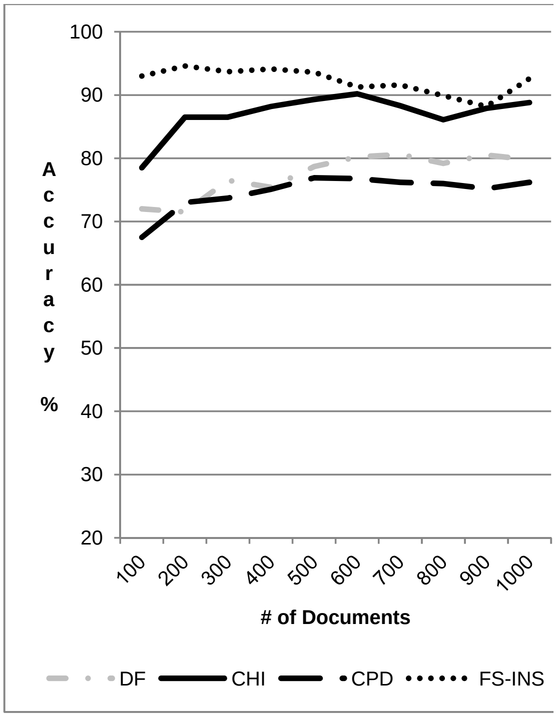

Fig. 5: The performance of OM with SVM as the classifier

Similar pattern was observed in opinion mining using k-NN classifier as shown in Fig. 4 and opinion mining using SVM classifier as shown in Fig. 5. The performance opinion mining with FS-INS as feature selection was better as compared to the other feature selection techniques. The performance of opinion mining with DF is the lowest since DF did not consider relationship of a particular feature with a class. It will sort the features based on the frequency of a particular features exist in both classes.

These experiments also showed that feature selection based on FS-INS was able to select the relevant features without any interference from the human beings. Features that exist with similar numbers in both classes are irrelevant and should not be selected even though they exist frequently in the corpus. Based on these experiments, the writer concluded that relevant features were features that exist in one class with a high frequency as compare to its existence in other classes and features that exist only in one class with very high frequency.

V. Conclusion

This paper discusses the use of a feature selection technique named FS-INS in opinion mining using online messages, which were created by the Malaysians. High number of noisy texts and improper uses of grammar and language syntax caused opinion mining with NLP techniques to be difficult. On the other hand, in machine leaning approach, a good feature selection technique was required. Inspired by the artificial immune system, a feature selection technique was developed in this study.

The experiments showed that the number of features selected by FS-INS was been reduced by 90% thus triggered the improvement of opinion mining's accuracy up to 15% in k-NN classifier and up to 6% in Naïve Bayes classifier. Incorporating this technique in real online application such as Customer Relation System or Frequently Question and Answering System (Q&A) is one of the future advancements that had been planned for future enhancement.

References

[1] 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

[2] 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

[3] Zheng, W., and Ye, Q. 'Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm', *Intelligent Information Technology Application, 2009.(IITA 2009),* 2009, pp. 335-338

[4] Zhai, Z., Xu, H., Li, J., and Jia, P. 'Feature Subsumption for Sentiment Classification in Multiple Languages', In *Proc. of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, Hyderabad, India, 2010, pp. 281-271

[5] Zhang, W., Yoshida,T. & Tang, X. 'A Study with Multi-Word Features in Text Classification', *In Proc. of the 51st Annual Meeting of the ISSS*, Tokyo, Japan, 2007 pp. 1-8

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

[7] 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

[8] Chen, L.S., and Chang, H. W. 'A Feature Selection Method for Classifying Textual Sentiment Data', downloaded from http://apiems.net/archive/apiems2010/pdf/AI/307.pdf , on January 1 2013.

[9] Hart, E., and Timmis, J. 'Application Areas of AIS: The Past, the Present and the Future', *Applied Soft Computing*, 2008, 8, (1), pp. 191-201

[10] Castro, L.N.d., and Timmis, J.I. 'Artificial Immune Systems as a Novel Soft Computing Paradigm', *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* 2003, 7, (8), pp. 526-544

[11] Dasgupta, D., and Niño, L.F. '*Immunological computation: theory and applications'* (CRC Press, 2008)

[12] Castro, L.D., and Timmis, J. '*Artificial immune systems: a new computational intelligence approach'* (Springler, 2002.)

[13] Secker, A., Freitas, A.A., and Timmis, J. 'AISEC: an artificial immune system for e-mail classification', In *Proc. of the Congress on Evolutionary Computation*, Canberra, 2003, pp. 131–139

[14] Oda, T. 'A Spam-Detecting Artificial Immune System', Master Thesis, Carleton University, 2005

[15] Guzella, T.S., Mota-Santos, T.A., Uchoa, J.Q., and Caminhas, W.M., 'Identification of SPAM Messages Using an Approach Inspired on the Immune System', *Biosystems,* 2008, 92, (3), pp. 215-225

[16] Secker, A., Freitas, A., and Timmis, J., 'AISIID: An Artificial Immune System for Interesting Information Discovery on the Web', *Applied Soft Computing,* 2008, 8, (2), pp. 885-905

[17] Zhang, Q., Luo, M., Wang, H., and Tan, J. 'A Medical Text Classification System Based on Immune Algorithm', In Proc. *Future Bio Medical Information Engineering, 2008. FBIE'08* 2008, pp. 433-436

[18] Zhang, Q., Luo, M., Xue, Y., and Tan, J.: 'Multi-class Text Categorization Based on Immune Algorithm', In. *Proc. of International Symposium on Intelligent Information Technology Application*, Shanghai, China, 2008, pp. 749-752

[19] Samsudin, N., Puteh, M., Hamdan, A.R., and Ahmad, M.Z.: 'Normalization of Common Noisy Terms in Malaysian Online Media', in *Proc. of Knowledge Management International Conference (KMICe)*, Johor Bahru Malaysia, 2012, pp 131-136.

[20] http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/

[21] Kwee, A., Tsai, F., and Tang, W., 'Sentence-Level Novelty Detection in English and Malay', in *Lecture Notes in Computer Science (LNCS)*, 2009, vol. 5476, pp. 40–51