Journal of Engineering and Applied Sciences 12 (13): 3534-3540, 2017
ISSN: 1816-949X

Enhanced Feature for Short Document Classification

Ali Abdulkadhim Hasan, Sabrina Tiun, Maryati Mohd Yusof,
Umi Asma' Mokhtar and Dian Indrayani Jambari
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Abstract: Now a days, the use of short text has been increased dramatically in which many applications are being relied on short text such as mobile messaging, breaking news social media and queries. The key challenging behind the short text lies on the limitation of acquiring context information from such text. This limitation increases both sparsity and ambiguity of the text. The traditional approaches that have been used for the classical text such as bag-of-words, seems to be insufficient due to the too limited information that could be extracted from the short text. This leads to loss the semantic knowledge and the semantic relations between the words within the short text. Hence, this study aims to propose a new feature selection method based on Interesting Term Count (ITC) with an external knowledge of WordNet and weighting to new weight (di) to identify the variation between classes on the base of ITC. The proposed feature selection approach aims at identifying the frequent terms without losing the semantic manner where the WordNet will be utilized in order to provide the semantic correspondences among the words within the short text. Furthermore, three classification methods have been used including support vector machine, J48 and Naive Bayes. The evaluation has been performed by applying the three classifiers with the proposed feature selection method and without the proposed feature selection method. Experimental results shown an outperformance of the classifiers with the proposed feature selection method. This can imply the effectiveness behind using the proposed ITC with external source knowledge for the short text classification.

Key words: Short text, text classification, feature selection, ITC, WordNet, NB, J48, SVM

INTRODUCTION

Recently with the dramatic expansion of the information, short text plays an essential role in terms of different applications such as web, mobile and other applications. The key difference behind short text lies on its sparsity in which the text is being described shortly and briefly compared to the classical text data (Sun, 2012). In this vein, the text will tend to be restricted to specific key words that could indicate the domain of such text. Apparently, this will increase the complexity of conducting specific operations on this text such as classification, categorization, clustering and extracting features.

Basically, short text could be found in multiple forms including blogs, social media, breaking news and queries. The nature of these forms of text is usually contain less length compared to the traditional text. The context is being summarized in order to reduce time consumption of reading. In this manner, the traditional approaches to represent the text such as the bag-of-words will be insufficient due to the shortness of the text where the information is too restricted. In addition, the short text sometimes tends to be non-formally written in which the syntax is not being considered. Obviously, this will hinder the process of utilizing syntactic approaches.

The challenges to utilize the traditional text task such as classification in the short text that, the short text does not provide adequate statistical information for effective similarity measure. Short text data further exacerbates the problem due to their sparse and noisy nature (Tang *et al*., 2012). Unlike short text, traditional documents are usually handled as vectors where the words or terms are represented as features whether one word (i.e., unigram) or multiple words (i.e., bigram or trigram). This makes the vector space of the traditional document classification is relatively high with tremendous set of features. In comparison, short text could not produce a high dimensional vector space where the length of the words is too limited.

Feature selection masterly established from a number of subgroup's features that is the most representative of the original feature set. It frequently reduced the implementation time in the text processing and increases the accuracy of classification because of removing some data outliers (Liu *et al*., 2010). In the short text domains,

Corresponding Author: Ali Abdulkadhim Hasan, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

precious feature selection is crucial to enhance the learning process effectively and efficiently. Short text extraction features traditional approaches and methods extracting a long text features have in common. The accuracy of the feature word is the key of text feature selection, according to segmentation and statistical accuracy. For the long and traditional text, there are many algorithms such NGram, vector space based methods models, statistical and improved algorithms (Kalchbrenner *et al*., 2014). Moreover, short text needs more study and design suitable methods depends on its specific characteristics.

Text classification suffer from high dimensionality as a result of large feature space. In addition, short text data further aggravates the problem because of their sparse and unstructured nature. Therefore, feature selection is an important step in improving the classification performance. However, existing feature selection methods cannot effectively extract these short text features and greatly reduce the classification performance of short text. Thus, it is necessary to provide more appropriate method of representation for the short text in order to improve the classification process of such text.

Short text has its own features such as shortness, sparseness and non-standard ability. Therefore, normal machine learning methods usually fail to achieve desire accuracy. However, short text classification is a challenging field because many technologies are in the initial stage as well as the difficulties of classification didn't get the excellent solution such as how to design dynamic short text stream. This study aims to enhance the feature representation for short text by utilizing ITC with an external knowledge source of WordNet. Such proposed features aim to enrich the short text by bringing more semantic correspondences in which the ITC will address the interesting counts.

Literature review: There are many studies that have addressed the problem of short text classification for instance, Wang *et al*. (2012) have proposed a new feature representation method for short text based on multiple features including strong thesaurus, Latent Dirichlet Allocation (LDA) and Information gain. Basically, thesaurus feature aims to provide semantic matches for the text's words whereas LDA aims to conduct a vector space for the terms in which the terms are being associated with corresponding documents.

In addition, Patra *et al*. (2012) have presented a classification method for short text produced by interviews with medical patients. The researchers have utilized the TF-IDF feature in which the unigram and bigram terms are being examined in terms of frequency. Furthermore a Senti WordNet knowledge source has been used in order to identify the emotion words. Finally, three classifiers have been used including Naive Bayes, Decision Tree and K-nearest neighbor.

Moreover, Li and Qu (2013) have proposed a classification method for short text using a feature selection approach called ITC. Basically, the researcher have clarified the limitation of TF-IDF compared to ITC. Consequentially, the researchers have enhanced the ITC by utilizing the two concepts of document distribution entropy and the position distribution weight. In fact such concepts are significantly contributed toward short text classification.

Zheng *et al*. (2014) have proposed a classification method for the short Chinese text using Chi-square feature and LDA feature. In fact, the proposed method will seek the contextual information from the text statistically.

Saif *et al*. (2014) studied the effect of removing a vital stop words to classify the sentiment of tweets through the application of traditional methods of feature selection. The researchers considered the five techniques: TF, TF1 (the words appear more than once), the IDF, term random samples and MI. It was based on the experimental evaluation in 5 sets of data on a small scale for Twitter, who belong to different areas. It has been selected for these two important classifiers: maximum entropy and Naive Bayes (NB). It has been to achieve the best results with the classification TF1and MI. A study submitted by Wang (2014) to extract features by improving the method of calculating the weight depended on words co-occurrence. Calculating the degree of co-occurrence identified as calculates the degree of relationship between each characteristic item and text. In fact, calculation method for the degree of co-occurrence is similar to the method of the calculation for conditional probability. The procedure is to find the location of the two words in a text which can reflect the co-occurrence and the compact degree of the two words. The 500 short text were chosen from tweet as dataset. The drawback of this method occurs in the location of the Characteristic Item, synonym and changed words. Zareapoor and Seeja (2015) used two techniques for feature selection and two techniques for feature extraction to enhance the performance of email classification. Email classification is difficult because of the high scattered dimensions features that affect the performance of mainstream works. Chi-Square and Information gain ratio used as feature selection techniques while Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) as feature extraction techniques. The study found that the performance of classification is better when using feature extraction.

Feature extraction methods (PCA, LSA) are not dependent on number of features chosen which an advantage in the text classification is since it is used to choose the correct number of features in the high dimensional space.

Furthermore, Mahajan *et al.* (2015) have proposed a feature selection approach based on Wavelet Packet Transform. Basically, the proposed WPT aims to identify the distance between the features in the vector space in order to distinguish the features. Experimental results shown superiority of the proposed WPT compared to the traditional techniques of feature selection.

Finally, Yin *et al.* (2015) have proposed a feature selection approach based on a semi-supervised learning method with Support Vector Machine (SVM). The proposed method has the ability to identify the significant features from the short text.

MATERIALS AND METHODS

The proposed method is composed of multiple phases as shown in Fig. 1. First phase aims to prepare the data in which multiple sub-tasks are being conducted such as tokenization, stop-word removing and Unigram. Second phase aims to utilize multiple features such as ITC and ITC with semantic knowledge of WordNet. Third phase is associated with the classifiers that are being used including NB, J48 and SVM. Final phase is associated with the evaluating the performance of the proposed method.

Tokenization: This phase aims to turn the sentences into series of tokens (i.e., terms). This process is vital in terms of handling the terms in the short text separately. This can be performed by utilizing the border of each word in order to separate the tokens.

Stop-word removal: Similar to the traditional text classification, short text contains several unwanted and unnecessary data. One of these data is the stop-words. Obviously, stop-words do not yield or affect the contextual information of the short text. Therefore, it is necessary to get rid of them. This can be performed by using a stop-words list in order to match all the candidates within the short text.

Unigram: This phase aims to handle the terms as a single unit called unigram. This is because the process of examining the frequency of each term is mainly relying on identifying unigram terms.

Feature selection: This phase aims to utilize multiple features from the short text. As mentioned earlier, the key challenging behind short text lies on the difficulty of extracting contextual information. Therefore, this study aims to proposed two set of features. The first set aim to utilize the ITC in order to identify the interesting pattern of frequency for the terms. ITC is considered to be a modified version of TF-IDF in which the term frequency is being replaced with a logarithmic adoption (Li and Qu, 2013). This process of replacement will avoid the limitation of TF-IDF in which the term frequency may mislead to unwanted class label. ITC can be computed as in Eq. 1:

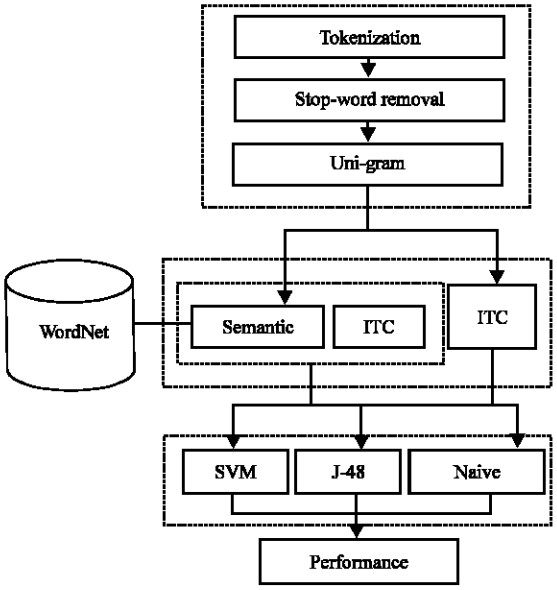

Fig. 1: Proposed method framework

$$Wid = \frac{\log(tf_{id}) \times \log\left(\frac{N}{n_i} + 0.01\right)}{\sqrt{\sum_{i=0}^{n} \log^2(tf_{id}) \times \log^2\left(\frac{N}{n_i} + 0.01\right)}} \quad (1)$$

Basically, Li and Qu (2013) have criticized the ITC in terms of the IDF where in some cases the terms are being given high value of IDF in accordance to specific class label, meanwhile this terms is irrelevant to this class. In this manner, the classifier will incorrectly classify such term to the irrelevant class label. Therefore, this study aims to utilize an enhanced version of ITC. The enhancement can be represented by adding the semantic aspect of WordNet. WordNet (Miller, 1995) is a large semantic knowledge source that contains tremendous kinds of semantic correspondences for each term such as synonyms, hypernyms and hyponyms. The process of enhancing ITC can be illustrated as follows.

For each term t in term set we calculate the semantic similarity between t and another term in feature according

to the Eq. 2. It assumes that the similarity between two concepts is the function of path length and depth in path-based measures (Meng *et al*., 2013):

$$Sim_{wp}(c1,c2) = \frac{2\times depth(lso(c1,c2))}{len(c1,c2)+2\times depth(lso(c1,c2))} \quad (2)$$

where, c1 and c2 are two different words are being extracted from the short text. When value of Sim_{wp} (c1, c2) is greater than a threshold of 0.8 we add the word to term's feature group, then count the terms in the group and assign to (S). Compute the average of the sum total of the number of the text appearing (t) and the number of text of terms of other text semantic similarity with t as follows:

$$\bar{n} = \frac{n_t + \sum_{k-1}^{s} n_t}{S+1} \quad (3)$$

where, n_t is the number of text containing term (t). From the Eq. 1 and 3 we can obtain the following Eq. 4:

$$W_{id} = \frac{\log(tf_{id})\times\log(\frac{N}{\bar{n}}+0.01)}{\sqrt{\sum_{i=0}^{n}\log^2(tf_{id})\times\log^2(\frac{N}{\bar{n}}+0.01)}} \quad (4)$$

Calculate the similarity based term frequency (stf) according to the following Eq. 5 (Howlett and Jain, 2005):

$$stf_{id} = tf_{id} \times \left(1+\frac{1}{dis(t_{id}, t\_stm_{id})}\right)^{tf\_sim} \quad (5)$$

Where:

t_{id} = The term in the specific text

tf_{id} = The term frequency of the term t in the document

t_sim_{id} = The similar term found to term t_{id}

tf_sim_{id} = The term frequency of the similar term t_sim_{id} in the document and dis(t_{id}, t_sim_{id}) is the distance between t_{id} and t_sim_{id}

To improve the distinguishing ability between classes a weight will be added as a factor to ITC for identifying which class is related to the text containing t as weight di = (1/(n-m+1). The (n-m) is the difference-value between the number of all text containing term (t) and the maximum number of text containing term (t) in certain class. Finally, the equation of enhanced ITC can be computed as follows:

$$W_{id} = \frac{\log(stf_{id})\times\log\left(\frac{N}{\frac{n_t+\sum_{k=i}^{s}n_t}{s+1}}+0.01\right)}{\sqrt{\sum_{i=0}^{n}\log^2(tf_{id})\times\log^2\left(\frac{N}{\frac{n_t+\sum_{k=i}^{s}n_t}{s+1}}+0.01\right)}}\times\left(\frac{1}{n-m+1}\right) \quad (6)$$

Hence, the enhanced ITC will be used to overcome the limitation of traditional ITC in terms of classifying the short text.

Classification: Once the features are being extracted using the enhanced ITC, three classifiers including NB, J48 and SVM are being used to classify the short text 1into their actual classes. The reason behind using three classifiers lies on the process of identifying the most compatible classifier with our proposed method.

Experimental results: In this study, the experimental results will be examined. This require identifying multiple aspects including dataset, evaluation method and the results. These aspects are being tackled in the following sub-sections.

Dataset: A benchmark dataset called search-snippets which is contains 12340 Web search snippets. This dataset has been used in many studies (Bouaziz *et al*., 2014; Chen *et al*., 2011; Phan *et al*., 2008; Sun, 2012; Xu *et al*., 2015). Such dataset contains multiple class labels (i.e., topics) including computers, culture-arts entertainment, education-science, engineering, health, politics-society and sport. This dataset was downloaded from the website (http://jwebpro.sourceforge.net/data web-snippets.tar.gz).

Evaluation method: Basically, the evaluation has been performed upon the testing set in which the training has been set to 60% and testing has been set to 40%. The evaluation has been conducted based on the common information retrieval metrics precision, recall, F-measure, TNR and FPR. Precision is the number of correctly classified instances in accordance to the total number of instances and it can be computed as:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Whereas Recall is the number of instances that have not been classified in accordance to the total number of instances and it can be calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

The overall accuracy which called F-measure based on Eq. 9:

$$F-\text{measure} = 2 \times \frac{P \times R}{P + R} \quad (9)$$

True Negative Rate (TNR) is the ratio of true negative predictions over the number of negative instances in the entire data set as shown in Eq. 10:

$$TNR = \frac{TN}{TN + FP} \quad (10)$$

False Positive Rate (FPR) is the ratio of the number of positive examples classified over the number of negative instances in the entire data set as in Eq. 11:

$$FPR = \frac{FP}{TN + FP} \quad (11)$$

RESULTS

This study aims to depict the results of the three classifiers with traditional ITC and with the proposed enhanced ITC. Table 1-7 show such results.

A comparative analysis will be presented in order to examine the performance results based on the use of similarity measure in order to improve the feature selection. Table 7 and Fig. 2 show the performance results by using ITC and Enhanced-ITC.

We can see that the accuracy result in Table 7 and TPR in Table 1 have improved using similarity measurement. The Wu-Palmer increased the accuracy from 85.6-88.1% as well as the TPR and F-measure have increased in the computers and culture-arts-entertainment classes.

The enhancement of proposed algorithm gives reasonable results with Naive Bayes classifier (in which 'reasonable' here is defined as a slight of improvement) where the accuracy has improved from 83.8-85.9% based on Wu-Palmer similarity measure. A TPR attains 4% increase in Busines and Education-Science classes, whereas in Engineering class, it increases just 2%. However, the F-measure has been amplified in these classes clearly. In addition, the worst performance of classification using J48 across Search-Snippets dataset, is the standard ITC which has gained accuracy of 79.4%. This rate is best improved by using ITC combined with Wu-Palmer similarity with accuracy up to 82.5%.

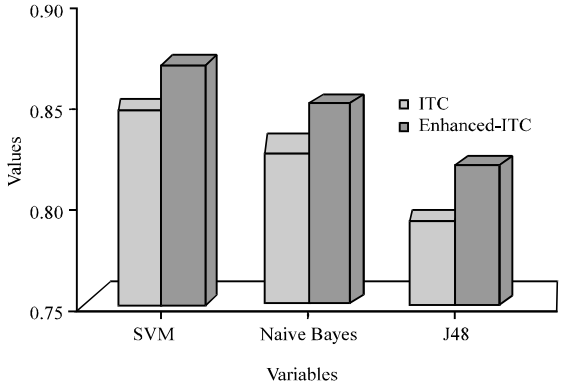

Fig. 2: Performance of ITC and Enhanced-ITC

Table 1: Precision and recall rate for SVM with ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.943	0.868	0.904	0.993	0.018
Computers	0.837	0.878	0.857	0.975	0.018
Culture-arts-entertainment	0.875	0.897	0.886	0.982	0.014
Education-science	0.912	0.795	0.849	0.989	0.028
Engineering	0.816	0.886	0.849	0.975	0.014
Health	0.854	0.833	0.843	0.978	0.025
Politics-society	0.870	0.816	0.842	0.978	0.033
Sports	0.767	0.892	0.825	0.965	0.014

Table 2: Precision and recall rate for SVM with enhanced ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.946	0.875	0.909	0.993	0.018
Computers	0.857	0.923	0.889	0.979	0.011
Culture-arts-entertainment	0.881	0.902	0.892	0.982	0.014
Education-science	0.919	0.850	0.883	0.989	0.021
Engineering	0.892	0.892	0.892	0.986	0.014
Health	0.854	0.833	0.843	0.978	0.025
Politics-society	0.864	0.864	0.864	0.978	0.022
Sports	0.850	0.919	0.883	0.979	0.011

Table 3: Precision and recall rate for NB with ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.943	0.868	0.904	0.993	0.018
Computers	0.837	0.878	0.857	0.975	0.018
Culture-arts-entertainment	0.875	0.897	0.886	0.982	0.014
Education-science	0.912	0.795	0.849	0.989	0.028
Engineering	0.795	0.886	0.838	0.972	0.014
Health	0.780	0.800	0.790	0.968	0.029
Politics-society	0.841	0.771	0.804	0.974	0.040
Sports	0.750	0.825	0.786	0.961	0.025

Table 4: Precision and recall rate for NB with enhanced ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.972	0.875	0.921	0.996	0.018
Computers	0.837	0.878	0.857	0.975	0.018
Culture-arts-entertainment	0.875	0.897	0.886	0.982	0.014
Education-science	0.943	0.805	0.868	0.993	0.028
Engineering	0.816	0.912	0.861	0.976	0.011
Health	0.800	0.889	0.842	0.972	0.014
Politics-society	0.884	0.809	0.844	0.982	0.032
Sports	0.778	0.833	0.805	0.964	0.025

Table 5: Precision and recall rate for J48 with ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.914	0.800	0.853	0.989	0.028
Computers	0.884	0.844	0.864	0.982	0.025
Culture-arts-entertainment	0.800	0.762	0.780	0.971	0.036
Education-science	0.829	0.674	0.744	0.978	0.049
Engineering	0.737	0.757	0.747	0.965	0.032
Health	0.732	0.811	0.769	0.961	0.025
Politics-society	0.800	0.818	0.809	0.967	0.029
Sports	0.674	0.906	0.773	0.951	0.011

Table 6: Precision and recall rate for NB with enhanced ITC

Class label	Recall	Precision	F-measure	TNR	FPR
Business	0.919	0.829	0.872	0.989	0.025
Computers	0.884	0.844	0.864	0.982	0.025
Culture-arts-entertainment	0.821	0.800	0.810	0.975	0.028
Education-science	0.865	0.762	0.810	0.982	0.035
Engineering	0.789	0.789	0.789	0.972	0.028
Health	0.800	0.865	0.831	0.972	0.018
Politics-society	0.795	0.814	0.805	0.968	0.029
Sports	0.738	0.912	0.816	0.962	0.011

Table 7: Accuracy performance results of ITC and enhanced-ITC

Classifier	ITC	Enhanced-ITC
SVM	0.856	0.881
Naive Bayes	0.838	0.859
J48	0.794	0.825

As shown in Table 7 using of SVM classifier with Wu-Palmer across Search-Snippets dataset produced higher accuracy of 88.1%. The worst performance is by using J48 with ITC resulted in an accuracy of 79.4%. The justification for such results is that the test set contained numbers of different concepts to learn with slightly large labels thus, the ability of SVM to learn can be independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data.

Basically, the vector space capability of SVM enables the proposed enhanced ITC to improve the classification accuracy. This can be represented by resembling more contextual information in the vector space. This is the reason that SVM shown superior results compared to the other classifiers.

Generally, these finding reflects the effectiveness of the proposed enhanced ITC in terms of classifying short text where the combination of ITC and the sematic weights mechanism has the ability to increase the contextual information which directly contributes toward improving the classification performance.

DISCUSSION

In fact to clarify the enhancement, it is necessary to accommodate a comparison with the state of the art approaches that examined the problem of short text classification. For example, Li and Qu (2013) have proposed a short text classification and obtained a precision of 88% and a recall of 82%. In addition, Wang *et al.* (2012) have proposed a short text classification method which gained a precision of 87% and a recall of 87%. Comparing these results with our proposed method's results of 93.8% for both precision and recall would significantly demonstrate that the proposed method has a competitive performance.

CONCLUSION

As a conclusion, we can see that the short text is shown to be difficult to be classified than long text which is related to larger data text. This can be explained mainly by saying that there are rare occurring in the words and thereby it will be hard to be captured semantically for such texts. The effect of term frequency is not adequate compared with the long documents. the challenges to utilize the traditional text task such as classification in the short text that, the short text does not provide adequate statistical information for effective similarity measure. Thus, short text data further exacerbates the problem due to their sparse and noisy nature. Moreover, short text data further aggravates the problem because of their sparse and unstructured nature. Therefore, feature selection is an important step in improving the classification performance.

This study has presented a new method for feature selection in terms of short text classification. Such method is an enhanced of ITC by combining a semantic weight produced by an external knowledge source of WordNet. Such combination would significantly increase the contextual information of the short text which directly improve the classification accuracy. Consequentially, three classifiers including NB, SVM and J48 have been used for the classification task. Results shown that the proposed enhanced ITC has outperformed the traditional ITC. In addition, SVM has shown the superior results compared to the other classifiers. This would demonstrate the effectiveness of combining the semantic weight for the ITC in terms of enriching the contextual information. For future researches, combining the distribution entropy and the position weight with the semantic approach and ITC would contribute toward enhancing the short text classification.

ACKNOWLEDGEMENT

This study is supported and funded by the University Kebangsaan Malaysia (UKM) under research grant: DIP-2016-033.

REFERENCES

Bouaziz, A., C. Dartigues-Pallez, C.D.C. Pereira, F. Precioso and P. Lloret, 2014. Short text classification using semantic random forest. Proceedings of the 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2014), September 2-4, 2014, Springer, Munich, Germany, pp: 288-299.

Chen, M., X. Jin and D. Shen, 2011. Short text classification improved by learning multi-granularity topics. Proceedings of the 22nd International Joint Conference on Artificial Intellig, July 16-22, 2011, AAAI Press, California, USA., pp: 1776-1781.

Howlett, R.J. and L.C. Jain, 2005. Knowledge Based Intelligent Information and Engineering Systems. Springer, Berlin, Germany,.

Kalchbrenner, N., E. Grefenstette and P. Blunsom, 2014. A convolutional neural network for modelling sentences. Master Thesis, Department of Computer Science, University of Oxford, Oxford, England.

Li, L. and S. Qu, 2013. Short text classification based on improved ITC. J. Comput. Commun., 1: 22-27.

Liu, Z., W. Yu, W. Chen, S. Wang and F. Wu, 2010. Short text feature selection for micro-blog mining. Proceedings of the IEEE 2010 International Conference on Computational Intelligence and Software Engineering (CISE), December 10-12, 2010, IEEE, Wuhan, China, ISBN:978-1-4244-5391-7, pp: 1-4.

Mahajan, A., S. Jat and S. Roy, 2015. Feature selection for short text classification using wavelet packet transform. Proceedings of the 19th Conference on Computational Language Learning, July 30-31, 2015, Association for Computational Linguistics, Beijing, China, pp: 321-326.

Meng, L., R. Huang and J. Gu, 2013. A review of semantic similarity measures in wordnet. Intl. J. Hybrid Inf. Technol., 6: 1-12.

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38: 39-41.

Patra, B.G., A. Kundu, D. Das and S. Bandyopadhyay, 2012. Classification of interviews a case study on cancer patients. Proceedings of 2nd Workshop on Sentiment Analysis where AI meets Psychology (COLING-2012), December 15, 2012, IIT Bombay, Mumbai, India, pp: 27-36.

Phan, X.H., L.M. Nguyen and S. Horiguchi, 2008. Learning to classify short and sparse text and web with hidden topics from large scale data collections. Proceedings of the 17th International Conference on World Wide Web, April 21-25, 2008, ACM, Beijing, China, ISBN:978-1-60558-085-2, pp: 91-100.

Saif, H., M. Fernandez, Y. He and H. Alani, 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014) Vol. 5, May 26-31, 2014, Curran Associates, Inc., Reykjavik, Iceland, ISBN:978-1-63266-621-5, pp: 1610-1617.

Sun, A., 2012. Short text classification using very few words. Proceedings of the 35th International ACM Sigir Conference on Research and Development in Information Retrieval, August 12-16, 2012, ACM, New York, USA., ISBN: 978-1-4503-1472-5, pp: 1145-1146.

Tang, J., X. Wang, H. Gao, X. Hu and H. Liu, 2012. Enriching short text representation in microblog for clustering. Front. Comput. Sci. China, 6: 88-101.

Wang, B.K., Y.F. Huang, W.X. Yang and X. Li, 2012. Short text classification based on strong feature thesaurus. J. Zhejiang Univ. Sci. C., 13: 649-659.

Wang, L.H., 2014. An Improved Method of Short Text Feature Extraction Based on Words Co-Occurrence. In: Applied Mechanics and Materials, Yarlagadda, P., C. Seung-Bok and K. Yun-Hae (Eds.). Trans Tech Publications, Switzerland, pp: 842-845.

Xu, J., W. Peng, T. Guanhua, X. Bo, Z. Jun and W. Fangyuan et al., 2015. Short text clustering via convolutional neural networks. Proceedings of the Conference on North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 31-June 5, 2015, IBM Watson, Denver, Colorado, USA., ISBN:978-1-941643-49-5, pp: 1-54.

Yin, C., J. Xiang, H. Zhang, Z. Yin and J. Wang, 2015. Short text classification algorithm based on semi-supervised learning and SVM. Intl. J. Multimedia Ubiquitous Eng., 10: 195-206.

Zareapoor, M. and K.R. Seeja, 2015. Feature extraction or feature selection for text classification: A case study on phishing email detection. Intl. J. Inf. Eng. Electron. Bus., 7: 60-65.

Zheng, C., D.K. Xiong and Q.Q. Liu, 2014. The short text classification method based on CHI feature selection and LDA topic model. Comput. Knowl. Technol., 14: 3182-3185.