Prediction of Individual Stock Movements in Bursa Malaysia using Online News

Pei Pei Chan and Kar Seng Loke[1]

[1] Monash University Malaysia, School of IT, Bandar Sunway,
46150 Petaling Jaya, Malaysia
Loke.Kar.Seng@infotech.monash.edu.my
http://www.infotech.monash.edu.my

Abstract. *Prediction results of individual stock movement in the Malaysian Stock Exchange using online news is presented. The technique combines automatic text extraction and neural networks prediction. The focus is on four companies listed in Malaysian stock market. Text preprocessing techniques are employed to process the news articles to produce phrase and word feature vectors. Backpropagation and Probabilistic neural network are employed as classifiers in the model and their performances are compared empirically. The findings show that the proposed approach gives prediction accuracy that is significantly better than random guessing. In addition, the empirical results also show that phrase input outperforms word input marginally in most prediction cases.*

1 Introduction

Stock markets are affected by the trends and activities in economic and financial industry. News related to these industries may play a role in affecting to a certain degree the movements of stock markets. There have been, indeed, numerous researches that examine the effects of news on financial markets, as cited in [1].

Research carried out to study the impact of news to Australian and New Zealand financial markets show that good news is associated with the increase in stock prices and exchange rates whereas bad news is linked to the contrary [1]. Goldberg and Leonard [2] also examined the influences of news contained in economic announcements on markets yields for United States and Germany. It was concluded that economic announcements certainly contain important news that could affect market yields. Correspondingly, economic announcements were regarded as an essential source of information for market participants.

Furthermore, Pui et al. [3] presented theories that stated that human behaviors are believed to be influenced significantly by news articles. Since human participants affect market development and news articles in turn influence the participants' behavior thus it implies that news articles may affect stock prices indirectly. The findings from these studies can serve as the basis for understanding how news significantly impact on stock markets. The abundance of online news articles provides valuable and potential resources for predicting the stock movements in the stock markets.

1.2 Research Overview

A prediction model that integrates text extraction and classification technique with a neural network classifier is proposed. This is an empirical research whereby experiments are carried out to evaluate the performance of the prediction model in stock movement prediction. Note, however, that the buy-sell signals, transaction costs, paper profits and other stock trading issues are not the subject of interest in this research thus will not be discussed.

In this research, neural networks are employed together with online news articles as input data in the prediction process. This paper presents the research on stock prediction that focuses on the listed companies in Malaysian stock market, which is called Bursa Malaysia. Four listed companies have been specially selected from two different sectors to be studied in detail. The two companies from the Finance sector are Public Bank Berhad and Malayan Banking Berhad. On the other hand, there are also Bandar Raya Developments Berhad and LBS Bina Group Berhad from the Property sector.

The first and main objective of this research is to examine the price movements of individual companies with online financial news. Four well-known companies listed in Bursa Malaysia are chosen in two different sectors. The empirical study seeks to predict the movements such as up, steady or down for the particular company on a certain day. This prediction will be done based on the news available online on that specific day. Online news is chosen because of ease

of processing and availability, and also of potential future automation. In any case, these news are essentially the same as what would be available on print.

2 Literature Review

We review some of the literature that uses textual data for stock movement forecasting. Most of the existing techniques discussed in this section are related to information retrieval and data mining.

2.1 Prediction

Wuthrich et al. [4] used data mining on online news articles to predict the daily movements of five major stock indices. Textual statements are claimed to affect the events in stock markets, i.e. certain stock price drops. One of the main features of this work is the use of priori domain knowledge, i.e., a set of 2-5 word keywords provided by a domain expert. Probabilistic rules are used to predict the trend of a particular index during the testing period. The system is reported to perform reasonably well in forecasting daily movements of the stock indices, and has managed to achieve an average accuracy of 43.6%.

News headlines and predefined expert chosen keywords are used as the input in forecasting intraday currency exchange rate movements in Peramunetilleke and Wong [5]. The system developed in this work aims at predicting the categorical outcome (i.e., up, steady, down) of the currency exchange rates based on the US dollar.

Three weight computation methods such as Boolean, TF x IDF (Term Frequency X Inverse Document Frequency) and TF x CDF (Term Frequency X Category Discrimination for Category Frequency) are discussed in the research. Based on experiments carried out using these three methods to compute keyword record weights, the study concluded that TF x CDF is the best performer among the three which achieve 51% accuracy.

NewsCATS (News Categorization and Trading System) is a stock price trend prediction system that is made up of three main components that process, categorize press release and generate trading strategies and recommendations [5]. Classification task is done using a Support Vector Machine based classifier [11].

The performance of the system is evaluated using market simulation to compare the average profit per trade against that of a random trader. The comparison results show that the system provides trading strategies that can achieve profit more effectively than randomly trading in the market.

Pui et al [3] proposed a framework for mining multiple time series concurrently and using textual document as source of prediction. The purpose is to investigate the inter-relationships between different stocks so as to determine the stocks that are influenced by a particular stock. The classifiers used in the experiments are based on the Support Vector Machine. Experimental results from the research show a significant increase in collective profit when mining multiple time series as compared to mining single time series.

In Sagar and Lee [7] articles from the Internet newsgroup is used with historical stock prices to predict the future stock prices. Natural Language Processing (NLP) is used to extract information from the text-based newsgroup articles. The initial investigation on the correlation between newsgroup articles and stock price movements found positive results.

2.2 Text classification and text preprocessing

Various techniques have been developed to build text classifiers for classification problems, for example, k-nearest neighbour [8], naïve Bayes [9][10], support vector machine [11] and neural network [12, 13].

Features extracted from the text documents mainly belong to two categories: bag-of-words and phrases [14, 15]. Many studies have been conducted to investigate the effects of these features on classification accuracy [16,17,18] however there seems to be no general consensus among these studies.

Generally, text classification process uses a vector representation of the documents to be classified. This vector representation contains terms from the documents as well as the weights assigned to each term. Various term weighting techniques, for example, term frequency (TF), inverse document frequency (IDF), TFxIDF [19, 20] have been developed.

Other text preprocessing techniques such as stemming [21,22], stop words removal [23,24] and principal component analysis [25] are commonly used in many text classification studies.

Numerous neural network approaches have been developed for text classification tasks. A 3-layer feedforward neural network that uses Backpropagation learning rule has been proven empirically to give good classification performance, which is measured by precision and recall [25]. The σ-FLNMAP neural model has been applied for classification of text documents from the Brown Corpus collection; it has been shown empirically to outperform other classification algorithms such as k-Nearest Neighbor and Naive Bayes classifier [12]. In another research, it was found experimentally that Backpropagation neural network outperforms a counterpropagation network in categorizing a set of 2,344 documents [26].

3. The Prediction Model

This section presents a model that incorporates neural networks to classify news articles for stock movement prediction. The networks used are Backpropagation and Probabilistic neural network that act as news classifier in the proposed model. News documents are initially required to undergo some preprocessing before being used as input to the classifier.

3.1 Prediction model based on classification

The prediction model proposed in this research adapts and incorporates text classification approach as part of the prediction process. Figure 1 gives an illustration of the prediction model. Text classification can be described as a task whereby a document from the document set is classified into a class from the set of predefined classes. In essence, the attributes of a text document determine to which predefined class the document will be assigned. As can be seen from the model in Figure 1, online news documents are used as input to the *Text Preprocessing* phase.

The final feature matrix representing the entire collection of news documents will subsequently be fed into the *Neural Classifier* which will strictly classify the documents into one of the three classes. The three classes are *Up*, *Steady* and *Down* whereby they are mutually exclusive.

Each news article is solely associated with a particular stock. The classification result for a news document is interpreted as a prediction based on the news content of that particular document. As indicated in Figure 1, historical closing prices together with reduced feature vectors are presented to the classifier as input-output pairs. The historical values are used for training the neural network. After training, the historical values are not used anymore in the prediction task.

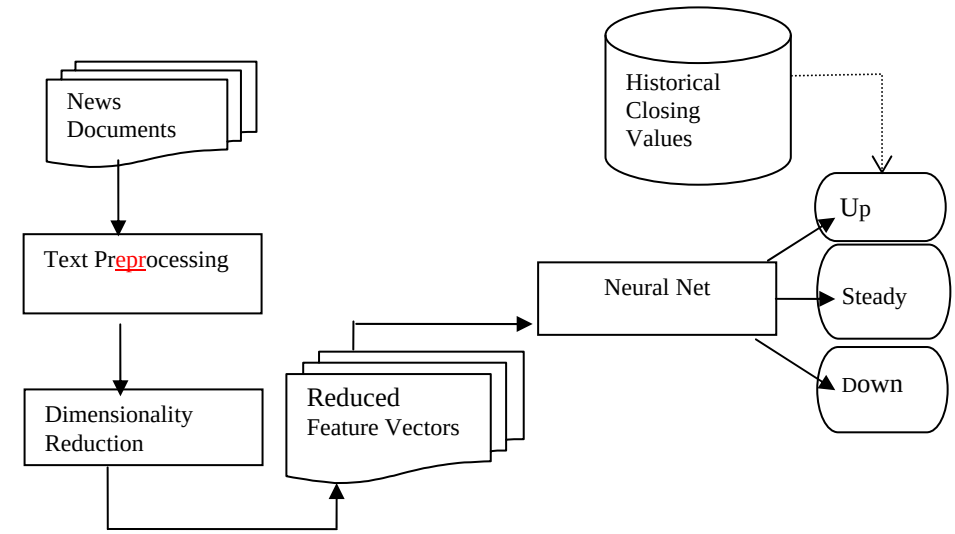

Figure 1: The model for news-based prediction

3.2 Text pre-processing techniques

The main purpose of text preprocessing is to identify the features that are representative of the news documents. More specifically, the textual content of the documents are translated into numeric representation.

Figure 2 shows the stages of text preprocessing and its processing sequence. The output is a set of feature vectors also known as document-feature matrix that represents all documents in the collection. This matrix is subsequently passed to principle component analysis (PCA) module where the dimensionality reduction process will begin.

After tokenization, the tokens tagged to indicate its part of speech using the MontyLingua algorithm that was adapted from transformation-based tagging algorithm [27]. Once all tokens have been tagged, regular expressions are employed to decide the tokens that will be grouped together as a phrase:

(RB | RBR | RBS | JJ | JJR | JJS)* (JJ | JJR | JJS)+

RB = Adverb
RBR = Adverb, comparative
RBS = Adverb, superlative
JJ = Adjective
JJR = Adjective, comparative
JJS = Adjective, superlative

Stop words removals are performed using 507 stop words from Lam and Savio [25]. Words that remained after all the stop words have been removed are to be stemmed using the Porter's stemming [27] algorithm.

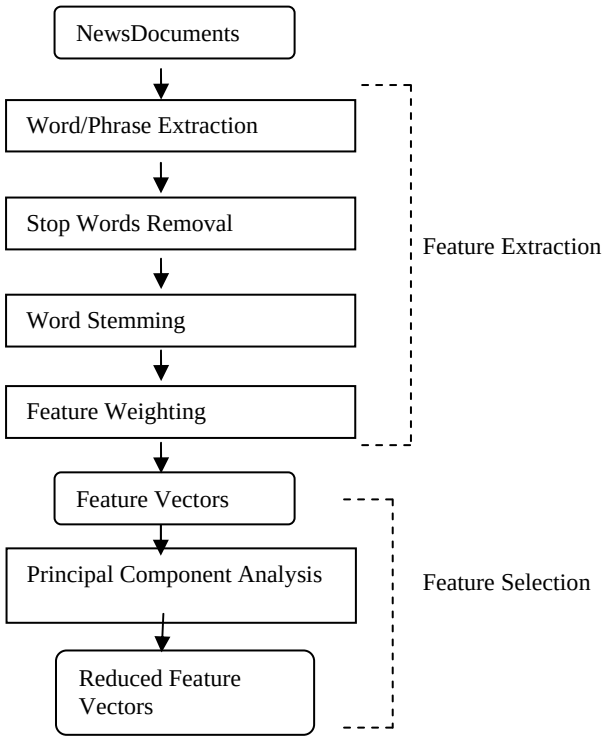

Figure 2: Flow of text preprocessing stages

After the stemming process, the duplicated stemmed features are removed. The *feature weight* captures the frequency count of the feature. As a result, features that appear frequently in a document will apparently have higher weight than those that appear less regularly.

A document-feature matrix, with its row and column as document and feature weight respectively, is used to represent the feature vectors for the whole collection of documents. PCA is then used to reduce the data dimension. Note, how-

ever, that this technique is not required when the classifier is PNN since this network has the ability to perform well with large feature vectors.

4 Performance Evaluations and Experimental Results

This section explains techniques for evaluating the performance of the prediction model and of the BPNN and PNN networks. Four main experiments are conducted to evaluate the performance of the prediction model and of the networks in predicting the stock movements for the four companies. The experimental results are presented and analyzed in this chapter.

4.1 Performance evaluation measure

The effectiveness of prediction is measured by its accuracy in classifying each news article to the correct class using:

$$\text{Accuracy} = \frac{\text{no. of correctly classified news articles}}{\text{total no. of all news articles}}$$

Each of the three class represents the different movements of the stock price thus there are three mutually exclusive classes namely up, steady and down.

4.2 Experimental Data

All the news articles are obtained from multiple sources such as AFX Asia (www.afxnews.com), Business Times (www.btimes.com.my), The Star Online: Business (biz.thestar.com.my) as well as Malaysian National News Agency: Bernama (www.bernama.com.my).

News articles collected for this research cover the period from 1st January 2003 to 31st October 2004, which is a total of one year and ten months. A total of 444 news articles for the four companies have been collected during this period of time. The period was chosen based on availability of the news data.

In the perspective of this research, the closing stock price is considered to have gone up only if it appreciates by at least 0.5% and above against the opening. Conversely, the price is regarded to have gone down when it drops by at least 0.5% or more.

4.3 Experimental Results

In Experiment 1 and 2, 2/3 of the news collection were used for training the classifier whereas the remaining 1/3 was for testing the classifier. These training and testing sets contain equally spaced points data picked from the original dataset. This is one way to select a potentially good subset of data that contains every possible example to train the network.

Table 1: Training and testing data distribution for the four companies

Company Name	Experiment 1 & 3		Experiment 4	
	Training Size	Testing Size	Training Size	Testing Size
Public Bank Berhad	105	53	127	31
Malayan Banking Berhad	102	51	123	30
LBS Bina Group Berhad	55	27	66	16
Bandar Raya Developments Berhad	34	17	41	10

A 50-trial run was carried out in each of the experiments that use BPNN as classifier and the average accuracy is computed. Due to the variation in output results from different experiments thus there is a need to run the experiments for a reasonable number of times to get the average result. Variation in results is attributed to the nature of the network whereby the initial weights and biases used in the network during training are obtained randomly.

On the other hand, unlike BPNN, PNN does not depend on random weights and biases thus the results from this classifier are expected to be relatively consistent. Consequently, in Experiment 2 the prediction process is conducted only once to produce prediction results.

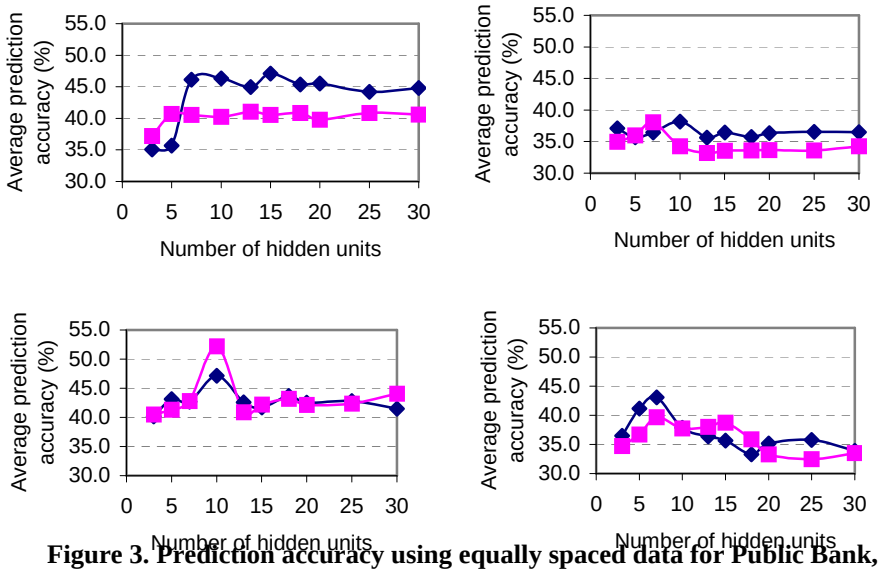

Figure 3. Prediction accuracy using equally spaced data for Public Bank, Malayan Banking, LBS Bina and Bandar Raya from left to right, top to bottom, in this order. Diamond indicates phrase input, and square indicates word input.

All the neural network experiments are conducted in the Neural Network Toolbox 4.0.1 of MATLAB Version 7. The BPNN uses the Levenberg-Marquardt learning algorithm with the network performance function (i.e., mean square error, MSE) set to 0.05.

4.3.1 Experiment 1.

In this experiment we use the BPNN with various hidden units. Figure 3 shows the average prediction accuracy. In this experiment, the inputs to the network are the equally spaced data taken from the whole collection of news articles. As shown in Figure 3, the phrase feature vector performs better than the word feature vector except in the case of LBS Bina. The graphs show that increasing the number of hidden units increases the average prediction accuracy.

Table 2 shows the prediction results with its corresponding p-values for each company. Each result is the highest accuracy obtained through *h*, where *h* denotes the number of hidden units. A p-value is the probability of observing a value of the test statistic as extreme as or more extreme than what is observed and it is calculated under the assumption that H_0 is true.

In the following, the probability of obtaining the observed outcome (i.e., the p-value) when the prediction is done based on random guessing is calculated. In this case, each prediction is independent from any other predictions. For a Binomial Distribution with parameters *n* and *p*, the mean = *np* and variance = *np(1-p)* where *p* is the probability of success and *n* is the number of times of prediction. If n is fairly large, the Binomial Distribution behaves approximately as Normal Distribution.

Since p-value ≤ 0.01 thus H_0 is rejected and it can be concluded that at the 1% significance level there is sufficient evidence to indicate that the data are consistent with H_1. In other words, the prediction accuracy is significantly better than random guessing. Overall, the prediction model using BPNN classifier is effective in predicting stock price movements for the four companies.

Company	Phrase			Word		
	h	Avg. accuracy (%)	p-value	h	Avg. accuracy (%)	p-value

Public Bank	15	47.09	≈0	13	41.06	≈0
Malayan Banking	10	38.16	1.0537E-07	7	38.08	1.6662E-07
LBS Bina	10	47.14	≈0	10	52.21	≈0
Bandar Raya	7	43.06	9.6535E-10	7	39.65	4.8279E-05

Table 2: p-value for the highest average prediction accuracy for each company

4.3.2 Experiment 2.

We test the performance of PNN using equally spaced data. Based on Table 3, in comparison with the 10% significance level all the p-value tests are significant except the test for Bandar Raya using phrase input. In most cases, the null hypothesis (H0: $\mu p = 0.3333$) still can be rejected at the 10% significance level. However, this test is rather weak as compared to Experiment 1 and 2 that use either 1% or 5% significance level.

Company	Test Data Size	Phrase		Word	
		Accuracy (%)	p-value	Accuracy (%)	p-value
Public Bank	53	47.17	0.0253	43.40	0.0814
Malayan Banking	51	45.10	0.0535	41.18	0.0924
LBS Bina	27	48.15	0.0790	48.15	0.0790
Bandar Raya	17	41.18	0.3261	52.94	0.0755

Table 3: Prediction accuracy using PNN

5 Conclusion

Results from Experiment 1 have shown that the proposed prediction model is competent in predicting stock price movements for the four individual companies. The aim of the first research objective in predicting stock price movements for individual companies has been achieved through the results from this experiment. All the prediction results are significantly better than random guessing. On top of that, the performance of BPNN classifier has also shown its effectiveness in the prediction task.

In Experiment 1, the results of phrase and word inputs are fairly consistent. There may be insufficient evidence to make a strong claim that phrase input performs better than word input. However, in most cases phrase input can be regarded as a marginally better performer in predicting the stock movements for the four companies.

Overall, results from the Experiment 1 and 2 have shown that the prediction model can perform significantly better than random guessing. Although the data size may be a bit small, the text preprocessing techniques employed are able to extract useful information from the data. As a result, the neural network in the prediction model is able to produce significantly good results for most of the prediction tasks.

References

1. Ellis, L. & Lewis, E.:The response of financial markets in Australia and New Zealand to news about the Asian crisis. RBA Research Discussion Papers rdp2001-03. Reserve Bank of Australia. (2001)
2. Goldberg, L. & Leonard, D.: What moves sovereign bond markets? The effects of economic news on U.S. and German yields. Current Issues in Economics and Finance, Vol. 9, No. 9, (2003) 1-7.
3. Pui, C. F. G., Xu, Y. J. & Wai, L.: Stock prediction: integrating text mining approach using real-time news. Proceedings of 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, Hong Kong. (2003) 395-402.
4. Wuthrich, B., Leung, S., Peramunetilleke, D., Lam, W., Cho, V. & Zhang, J.: Daily Stock Market Predictions from World Wide Web Data. Int'l Conference KDD98. (1998) 269-274.
5. Peramunetilleke, D. & Wong, R. K.: Currency exchange rate forecasting from news headlines. Proceedings of the Thirteenth Australasian Conference on Database Technologies, Melbourne, Australia, Australian Computer Society, Inc., Darlinghurst, Australia, vol. 5. (2002) 131-139.
6. Mittermayer, M.-A.: Forecasting intraday stock price trends with text mining techniques. Proceedings of the 37th HICSS'04, Big Island, Hawaii, 5-8 January 2004, IEEE Computer Society. (2004) 64-73.

7. Sagar, V. K. & Lee, C. K.:A neural stock price predictor using qualitative and quantitative data. Proceedings of the 6th ICONIP'99, Perth, W.A., Australia, 16-20 November 1999, vol. 2. (1999) 831-835.
8. Yang, Y.: An evaluation of statistical approaches to text categorization. Technical Report CMU-CS-97-127, Carnegie Mellon University, April (1997)
9. McCallum, A. & Nigam, K.:A comparison of event models for Naïve Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. (1998) 41-48.
10. Sahami, M.: Learning limited dependence Bayesian classifier. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press. (1996) 335-338.
11. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Springer, Heidelberg. (1998) 137-142.
12. Petridis, V. Kaburlasos, V. G., Fragkou, P. & Kehagias, A.: Text classification using the σ-FLNMAP neural network. Proceedings of the IJCNN '2001, Washington D.C., 14-19 July 2001, vol. 2. (2001) 1362-1367.
13. Yang, Y. & Liu, X.: A re-examination of text categorization methods. 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, USA. (1999) 42-49.
14. Liao, C, Alpha, S & Dixon, P.:Feature preparation in text categorization. Australasian Data Mining Conference (AusDM 03), Canberra, Australia, 8 December. (2003)
15. Lewis, D. D.:Feature selection and feature extraction for text categorization. Proceedings of Speech and Natural Language Workshop, Harriman, New York, 23-26 February, Morgan Kaufmann, San Mateo, C.A.(1992) 212-217
16. Scott, S. & Matwin, S.: Feature engineering for text classification. Proceedings of 16th ICML-99, Morgan Kaufmann, San Francisco, US. (1999) 379-388.
17. Mladenic, D. & Grobelnik, M. D.: Word sequences as features in text-learning. Proceedings of the 17th ERK'98, Ljubljana, Slovenia. (1998) 145-148.
18. Dumais, S., Platt, J., Heckerman, D. & Sahami, M.: Inductive learning algorithms and representations for text categorization. Proceedings of the 7th ACM Int'l CIKM98, Bethesda, US, ACM Press, New York, US. (1998) 148-145.
19. Salton, G. & Buckley, C.:Term-weighting approaches in automatic text retrieval. Information Processing and Management, vol. 24, no. 5. (1998) 513-523.
20. Cho, V., Wuthrich, B., & Zhang, J.: Text processing for classification. In: Special issue on Financial Analysis using Distributed Data Mining, Journal of Computational Intelligence in Finance, vol. 7, no. 2. (1999) 6-22
21. Hull, D. A.: Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, vol. 47, no. 1. (1996) 70-84.
22. Xu, J. & Croft, W.: Corpus-based stemming using co-occurrence of word variants. ACM Transactions on Information Systems, vol. 16, no. 1, TR96-67. (1998) 61-81.
23. Yang, Y. & Liu, X.: A re-examination of text categorization methods. 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, USA. (1999) 42-49.
24. Riloff, E.: Little words can make a big difference for text classification. In: Edward A. F., Ingwersen, P. and Fidel, R. (eds.): Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, USA, ACM Press, New York, USA. (1995) 130-136.
25. Lam, L. Y. & Savio, D.: Learned text categorization by backpropagation neural network, Master thesis, The Hong Kong University of Science and Technology. (1996)
26. Ruiz, M. E. & Srinivasan, P.: Automatic text categorization using neural network: Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, Medford, New Jersey. (1998) 59-72.
27. Porter, M. F.:An algorithm for suffix stripping. Automated Library and Information Systems, vol. 14, no. 3. (1980) 130-137.
28. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, vol. 21, no. 4. (1995) 543-566.