Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014
ISSN 2074-8523; Copyright © SCRG Publication, 2014

Text Opinion Mining to Analyze News for Stock Market Prediction

Yoosin Kim[1], Seung Ryul Jeong[1], Imran Ghani[2]

[1]Business IT Graduate School, Kookmin University, Seoul, South Korea
e-mail: (trust,srjeong)@kookmin.ac.kr
[2]Universiti Teknologi Malaysia (UTM), *Skudai, Johor Bahru, Johor, Malaysia*
e-mail: imran@utm.my
*Corresponding author: Seung Ryul Jeong

Abstract

This is a known fact that news and stock prices are closely related and news usually has a great influence on stock market investment. There have been many researches aimed at identifying that relationship or predicting stock market movements using news analysis. Recently, massive news tests, called unstructured big-data, have been used to predict stock price. In this paper, we introduce a method of mining text opinions to analyze Korean language news in order to predict rises and falls on the KOSPI (Korea Composite Stock Price Index). Our method consists of carrying out the NLP (Natural Language Processing) of news, describing its features, categorizing and extracting the sentiments and opinions expressed by the writers. The method then identifies the correlation between news and stock market fluctuations. In our experiment, we show that our method can be used to understand unstructured big-data, and we also reveal that news' sentiment can be used in predicting stock price fluctuations, whether up or down. The algorithm extracted experiments can be used to make predictions about stock market movements.

Keywords: *Big Data, Text Mining, Opinion Mining, Stock Prediction, KOSPI, NLP.*

1 Introduction

Recently, unstructured text data, also known as big data, in web, mobile and SNS communications has rapidly emerged as a great source of useful information, ranging from news articles to personal opinions. Accordingly, text mining technology is becoming increasingly important as a means of rapidly analyzing

massive texts in order to derive significant information intelligently. It has many potential applications that could be used very effectively in such sales, marketing, manufacturing and various other areas [1].

In particular, the prediction of stock values using the news or personal opinions on the web constitutes an attempt to extract meaningful characteristics from texts using text mining; classify the news as good or bad for stock prices, simulate investment and predict price trends. However, it is still difficult to analyze and extract meaning because such text data comes in many diverse types, and comprises multi-meanings, complex words and various features.

The news, generally, presents both positive and negative aspects of the stock markets in a somewhat neutral tone, making it difficult to identify the underlying truth behind such news. Furthermore, there is a danger that the news may be analyzed and interpreted differently depending on the writer [5, 10, 12].

Thus, in order to overcome such limitations, we propose an text opinion mining system for supporting decision-making to invest, , whereby massive news articles, unstructured big data, are gathered, parsed, tagged, analyzed, and converted to opinions suitable for making stock market predictions. Also, we built a stock market oriented sentimental word dictionary; a lexical resource for sentiment analysis and opinion mining, as opposed to a general purpose a sentimental word dictionary. At the end, we experimented with news originating from two different media, and tested its accuracy in forecasting stock price fluctuations.

2 Related Works

2.1 Opinion Mining

Opinion mining, as a sub-discipline with data mining and computational linguistics, is referred to as the computational techniques used to extract, classify, understand, and assess the opinions expressed in various online news sources, social media comments, and other user-generated content [3]. 'Sentiment' analysis is often used in opinion mining to identify sentiments, affect, subjectivity, and other emotional states in the online text.

Works on opinion mining, such as that by [6], showed the effectiveness of automatic movie review mining and summarization of movies. The [13] proposed a technique by which to list ranked product reviews according to the intentions of the searcher. The [4] showed how market moods monitored from Twitter may be used to predict the flow of stock prices. Many works have focused on predicting stock prices using news opinion mining [7, 8, 9, 10, 11].

In order to more accurately extract opinions and sentiments from a text, it is very important to build up a lexicon of opinion mining. If a lexicon such as a sentimental word dictionary is developed properly, the results of opinion mining

will be good. Furthermore, sentimental word dictionaries are more effective when domain specific characteristics are taken into consideration [12].

2.2 Stock Prediction Using the News

Many works over the years have continued to prove that the news is closely related to stock prices. In particular, with the recent explosive increase in the amount of unstructured text data from the internet, mobile channels, and SNS (Social Network service), there have been attempts to predict stock movements using such text data. [7], using the stock prediction system NewsCASTS (i.e. the News Categorization And Trading System, which consists of three engines, namely, news pre-treatment, categorization and trading), analyzed media news on specific companies, and experimented with a comparison between news and stock price flows. The [10] gathered, analyzed and extracted individual investors' opinions disclosed on the web, analyzed their sentiments, calculated their author's reliability, and predicted the stock values of three companies via machine learning. The [8] proposed the machine learning system, AZFinText(Arizona Financial Text System), to infer stock price prediction variables from the news, and showed higher returns than the market average through trading simulation.

In Korea, [11] proposed an automatic news classifier and showed that the use of pattern matching-based news classification delivers a rate of accuracy of 69% when predicting whether stock prices will rise and a rate of 64% when predicting whether they will fall. The [9] extracted the features of news texts to compare stock price variations, and experimented with a classifier to predict whether specific companies' stock prices would go up or down, using the Naïve Bayesian model. The [5] suggested an intelligent investment model by text opinion mining which analyzes the sentiments of 'news big data', and showed that predictions using a logistic regression analysis achieved a rate of accuracy of 70.0% for increases in stock prices and 78.8% for decreases.

3 Approach

3.1 System Overview

The Figure 1 shows the overall outline of our system. The first step consists of news gathering. In this step, we got a scrapping of the online economic news boards on Naver.com and stored the data in a database. The next step consisted of natural language processing in order to extract sentiment from unstructured news texts. We removed stop words such as punctuation, numbers, English, html tags, and so on. The remaining useful words were then used to build a dictionary of sentiments. We conducted a sentiment analysis and opinion mining using the sentiment word dictionary, and then conducted supervised learning experiments aimed at predicting rises and falls in stock prices.

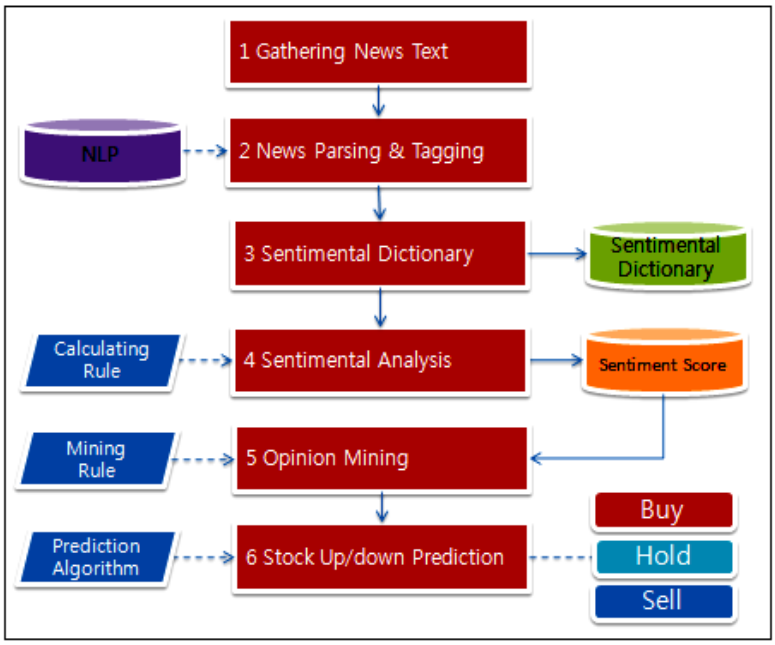

Fig.1: System overview

3.2 Data Collection

For the experiment, we gathered 78,216 economic news articles of media company M and H, over a period of one year (2011) from Naver.com (the No.1 portal in Korea), using scrapping technology. The two media display some different characteristics in terms of news. The online media M, a new entity established in 1999, focuses more on stock investment issues, and the rapid delivery of information on market situations and investment strategies. On the other hand, H media's vision is represented as 'the global comprehensive economic media'. H media, launched in 1964, provides a wide range of economic information, realty and industrial issues encompassing newspapers, TV, and the Internet.

The experiment divided the data into seven months of news articles from January 1 to July 31, 2011 as a learning data set, and into five months of news articles from August 1 to December 31, 2011 as a verification data set. Of the data, the number of media M's news articles amounted to 44,305, while that of media H's news' articles amounted to 33,911. The Table 1 shows the data sets by media.

Table 1: Experiment Data Set

Data Set	M	H	M+H
Training Data(1/1~7/31)	25,955	16,023	41,978
Test Data(8/1~12/31)	18,350	17,888	36,238
Total	44,305	33,911	78,216

3.3 Sentiment Analysis

Analyzing the sentiment of the news involves recognizing and defining the 'emotional state' expressed in the text. The sentiment word dictionary plays a crucial role in opinion mining to build linguistic resources, which classify sentiment polarity, quantify the breadth of sentiment, and discriminate between sentiments. In particular, [12] built a stock domain specific dictionary, which showed greater accuracy compared with general sentiment dictionaries in terms of its ability to predict price stock movements. Similarly, we extracted sentiment word from the news using NLP, calculated the sentiment score, and mining the opinion. The Figure 2 shows a flow diagram of the development of a sentiment word dictionary aimed at predicting rises and falls in stock prices.

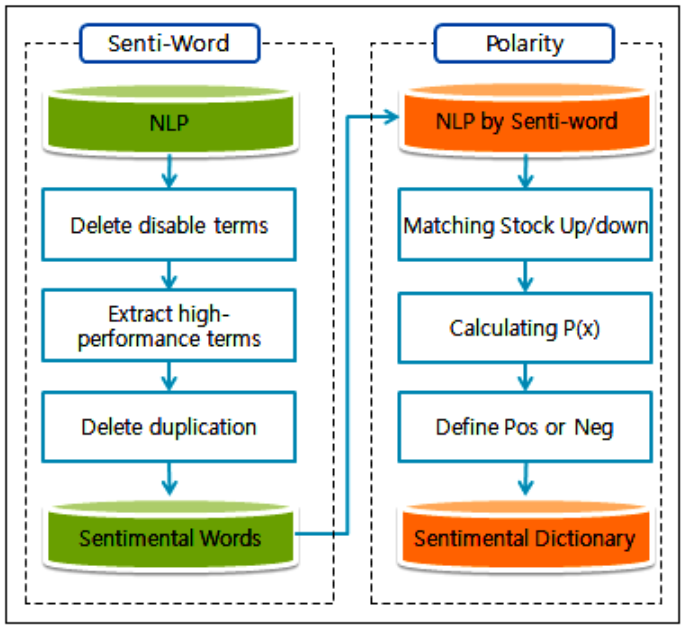

Fig.2: Flow of development of dictionary

In more detail, to develop a sentiment word dictionary, two phases are needed. The first phase consists in deleting stop words such as punctuation, numbers, English and one-character words, extracting high performance words using term frequency, and cleaning duplicate words and proper noun as product name.

The next phase consists in forming an opinion of the news by calculating the probability of recurrence of certain 'sentimental' words. The sentiment of a word is defined as the ratio of the number of stock price up or down following to the total number of news containing the word, whose calculation formula is expressed as follows:

$$
\begin{aligned}
&\bullet\ Word(i,j) = \begin{cases} 1 \begin{pmatrix} Doc(j)\ including\ Word(i), \\ Doc(j)\ next\ day\ stock\ price\ Up \end{pmatrix} \\ 0(the\ rest) \end{cases} \\
&\bullet\ Word(i).NumDocs = the\ total\ number\ of\ news\ include\ Word(i) \\
&\bullet\ Word(i).SentiScore = \frac{\sum_{i=1}^{n} Word(i,j)}{Word(i).NumDocs}
\end{aligned} \tag{1}
$$

In the formula, the score for a sentimental word ranges from a maximum of 1 to a minimum of 0. In other words, 1 is a fully positive sentiment whereas 0 is an entirely negative one. The sentiment score of a news article is calculated based on the average sentiment score for all sentimental words contained in the news. Likewise, the sentiment score for a day is calculated by the average sentiment score of the total news of that day.

3.4 Stock Predictions

The sentiment score should be changed an opinion to predict a rise or fall in the stock price. An opinion for prediction is decided to be positive, or not on a certain point. A next trading day's stock price is predicted to rise, or, conversely, predicted to fall. Likewise, since the setting of threshold criteria for predicting stock price fluctuation is crucial. The Figure 3 shows a flow diagram of the learning experiment for predicting stock prices. The ratio of accurate predictions to total predictions is defined as 'prediction accuracy'.

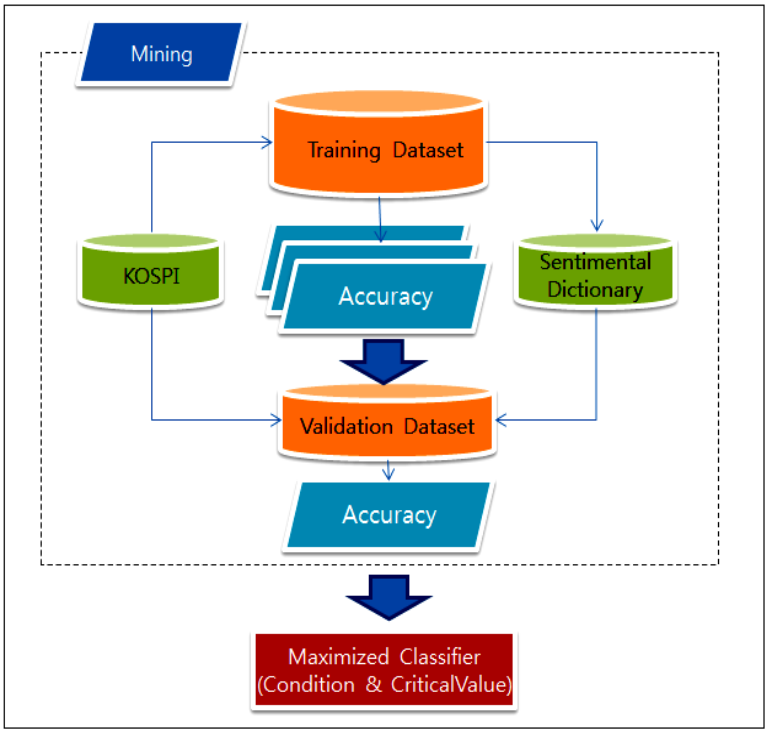

Fig.3: Experiment model overview

In addition, we tried to compare the news offered by two different media, and measured the prediction accuracy of each media at the threshold.

4 Experiments and Results

4.1 Evaluation

The accuracy of opinion mining in new sentiment analysis can be evaluated using statistical measures such as recall, precision, F1-score [7, 20, 11]. Accuracy, which is defined as the percentage of sentiments correctly predicted, is one method of evaluating. The quality of the results is also measured by comparing two standard performance measures, namely, recall and precision. Recall is defined as the proportion of positive sentiments which are correctly identified.

$$Recall = \frac{Positive\ instances\ predicted}{Total\ positive\ instances} \tag{2}$$

Precision is defined as the ratio between the numbers of correct sentiments predicted to the total number of matches predicted:

$$Precision = \frac{True\ positive\ instances\ predicted}{Total\ instances\ predicted} \tag{3}$$

But this would often decrease the precision of the result. In general, there is an inverse relationship between recall and precision. An ideal learning model would have both high recall and high precision. Sometimes, recall and precision are combined together into a single number called F1, which represents a harmonic mean of recall and precision:

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

4.2 Analyzing Sentiment

In previous works, the use of a domain specific dictionary rather than a general dictionary when analyzing the sentiment of news using a sentimental word dictionary showed greater accuracy of prediction [12]. That being the case, we developed a stock market specific dictionary and evaluated the sentiment of the words appearing in the news. In order to develop a stock domain specific dictionary, we performed NLP of 78,216 news articles. At the first parsing, we had over ten million words, but, after completing several pre-processing and selection procedures, we finally chose seven hundred sentimental words. The

selected words were calculated by a sentiment scoring formula. The Table 2 is a part of the sentimental word dictionary.

Table 2: Sample of sentimental words

Neg-Word	Score	Pos-Word	Score
더블딥(double dip)	0.167	경기부양(pump-priming)	0.700
동결(freeze)	0.246	신뢰(confidence)	0.692
방어(defense)	0.296	확충(expand)	0.677
급락장(slump)	0.333	호전(improve)	0.647
비교(compare)	0.341	정상(peak)	0.625
근거(basis)	0.365	합의(agree)	0.612
요구(require)	0.365	대장주(leading stock)	0.608
기준금리(base rate)	0.365	강한(strong)	0.565
우려감(fear)	0.366	사자(buy)	0.563
낙폭(range of drop)	0.382	급등세(sudden rise)	0.563

If the sentiment score of a word is closer to 1.0, it means that the word is more positive. Conversely, if a word is nearer to 0.0, it means the word is negative. As a result, some rare words showed extremely negative or positive polarity, while most words fell in the middle of the range, i.e. 0.3 ~ 0.7. Also, regarding the sentiment of the words, there was a relatively equal distribution of both positive and negative words. NewsOpn, i.e. the sentiment scores of news, was also calculated using a sentiment scoring formula. DayOpn was same way. The Table 3 shows some of the results of the calculation of news and day opinions.

Table 3: Sample of sentiment scores

DATE	NewsID	NewsOpn	DATE	DayOpn
2011-01-04	640	0.3000	2011-01-13	0.1735
2011-01-04	652	0.1765	2011-01-14	0.1521
2011-01-04	661	0.1600	2011-01-15	0.1801
2011-01-04	663	0.1053	2011-01-16	0.1366
2011-01-04	667	0.3103	2011-01-17	0.1636
2011-01-04	668	0.2424	2011-01-18	0.1808
2011-01-04	669	0.1429	2011-01-19	0.1714
~	~	~	~	~
2011-07-30	43011	0.1500	2011-07-22	0.2022

2011-07-30	43012	0.2353	2011-07-23	0.1250
2011-07-30	43013	0.0000	2011-07-24	0.1772
2011-07-31	43015	0.2319	2011-07-25	0.1905
2011-07-31	43016	0.1951	2011-08-26	0.2043
2011-07-31	43016	0.1860	2011-07-27	0.2334

DayOpn, Sentiment score of a day, was converted into stock price fluctuation prediction variables. We tried to determine the threshold, the critical value with the highest prediction accuracy by test.

4.3 Predicting Increases/Decreases in Stock Price

The aim of the experiments is to determine whether there is a correlation between the news sentiment and the rise of fall of stock prices, and whether the results of prediction differ according to each media. For the experiment, we separated three groups, media H, M and H + M, from the data set. The Figure 4 shows the level of prediction accuracy by the critical value of each type in the training data set.

In Figure 4, media M news' opinions showed the best prediction accuracy of 65.2% at the critical value of 0.22. H news' opinions, meanwhile, showed 60.3% prediction accuracy at the critical value of 0.19. The difference in accuracy between the two media is 5%, which means that M's news, compared to H's news, could predict with greater accuracy any rises and falls in stock prices. Actually, in terms of the content of the news articles, media M presents stock market situations, prospects and so forth in a comparatively clear tone, while media H presents more macroeconomic infrastructure and post-incident evaluations.

Another interpretation of the difference is that the newer media (M) is an online exclusive channel that focuses aggressively on stock market news coverage, whereas the older player (M) focuses on the overall economic sector rather than on presenting a specific area and predictions. On the other hand, M+H news' opinions showed 60.1% prediction accuracy at the critical value of 0.11. It is understood that the mixed analysis of news articles of the two somewhat contrasting media made the classification of opinions unclear, thereby lowering accuracy.

For the verification, thresholds drawn from the training data sets were applied to the test data sets, and the prediction accuracy of each group was tested. The Figure 4 shows the results of the experiment for the three groups in the test data set. Media M news' opinions, which showed the highest prediction accuracy in the training data set, also showed the highest accuracy of 54.8% in the test data

set. Media H news' opinions showed a rate of accuracy of 52.3%, a lower figure than that obtained by the training data set, while M+H news showed a very low level of accuracy at just 48.1%.

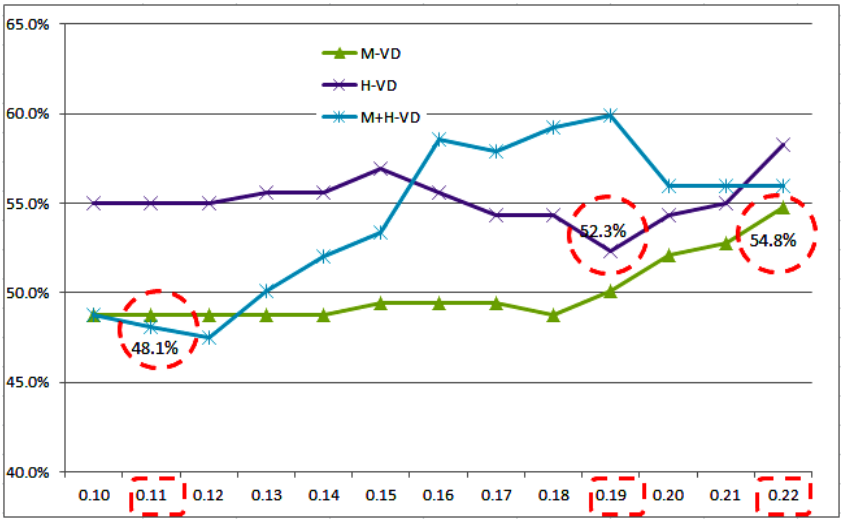

Fig.4: Threshold and accuracy in the test data

While the above results are focused on determining the threshold and prediction accuracy through the training and test data sets, the next thing that needs to be examined is how the prediction was performed to assess the quality of prediction. In order to assess the prediction quality, as mentioned in the evaluation, we conducted F1, the harmonic mean of recall and precision. Generally, if the prediction accuracy is higher, the result is better, and, when the level of accuracy is identical, if the F1 score is higher, it shows a higher degree of predictive power.

In Table 4 each experimental group's prediction accuracy and F1 featured in the training data sets were evaluated as follows: media M news' accuracy was the highest, at 0.652, compared to the other experimental groups, and its F1 was also the best at 0.626. This must mean that media M's news has the best prediction probability.

Table 4: Prediction quality in the training data

Training Data	Max Critical Point	Accuracy α	*F1*Score
M	0.22	0.652	0.626
H	0.19	0.603	0.526
M+H	0.11	0.594	0.736

Then, we tested the prediction accuracy and quality of the test data-which inherited threshold of the training data set. The result is shown in Table 5 where media M's news still displayed the highest prediction power. Although its

accuracy was 0.10 points lower compared to the training data, it showed prediction accuracy of 0.550 and an F1 of 0.630. This result means that rises and falls in stock prices can be sufficiently predicted using news opinion mining.

Table 5: Prediction quality in the testing data

Training Data	Max Critical Point	Accuracy α	*F1*Score
M	0.22	0.550	0.630
H	0.19	0.532	0.532
M+H	0.11	0.484	0.652

Ultimately, it showed that, although there are differences in the level of prediction accuracy and quality according to each media's opinion falls and rises in stock prices could be predicted using news opinions; and that sentiment analysis and the threshold, as learned through opinion mining, could be effectively used to predict the actual stock market.

In summary, news articles have their own characteristics, depending on the media. We showed that news articles are classified, opinions are extracted from there, an appropriate critical value, threshold, is performed, and then stock up/down is predicted. However, we should be aware that media's characteristics can render the result of opinion mining somewhat unclear and lower its predictive power.

5 Conclusion

This study, assuming that news and stock prices have a close correlation, sought to find patterns in the new that could be useful in predicting positive and negative fluctuations in stock prices. Many previous studies revealed that the news influences stock prices, and a number of studies on stock price prediction have been made using actual news articles. However, we have conducted a novel attempt to compile a stock domain specific sentimental word dictionary from the news as unstructured big data, to analyze 'sentiment' using that dictionary, and to mine opinions in order to predict stock price fluctuations (i.e. up/down movements).

As a result, we built a stock domain specific dictionary via the NLP of 78,216 news articles take from two different style media and showed the sentimental words, news sentiments and opinions calculated using the dictionary. Then, by conducting a stock prediction experiment, we found that the opinions extracted from the news could be useful in accurately predicting stock market movements and F1 features. Furthermore, we recognized that the media have their own characteristics, so the accuracy of predicting stock market price movements could differ depending on the media.

However, there are still many areas in this study which could be investigated in more detail and extended. The news sample of just one year used in this study may not be large enough, and the use of two media, M and H, may also be insufficient. It is particularly difficult to carry out NLP of Korean language when building a sentimental word dictionary. The Korean language varies greatly in its usages, and has various contextual meaning, and homonyms. Therefore, future studies will have to gather longer periods of big data, analyze more diverse media outlets, and reflect Korean linguistic characteristics more carefully. Furthermore, if attributes such as market prospects, overseas news, and corporate performance are classified, it will be possible to analyze its effects on stock prices and the predictive power according to news types more closely.

References

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, Vol.2, Nos.1–2, 2008, pp.1–135.

[2] Fu, K. Lee, D. Sze, F. Chung and C. Ng, "Discovering the Correlation between Stock Time Series and Financial News," *Proceedings of the 2008 IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008, pp.880-883.

[3] H. Chen and D. Zimbra, "AI and Opinion mining," *IEEE Intelligent Systems*, May/June 2010, pp.74-80.

[4] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, Vol.2, 2011, pp.1–8

[5] Kim, Y., N. Kim, S.R. Jeong, "Stock-index Invest Model Using News Big Data Opinion Mining," *Journal of Intelligence and Information Systems*, Vol.18, No.2, 2012.6, pp.143-156.

[6] L. Zhuang, F. Jing and XY. Zhu, "Movie Review Mining and Summarization," *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, November 2006, pp.43-50.

[7] M. A. Mittermayer and G. F. Knolmayer, "NewsCATS: A News Categorization And Trading System," *Proceedings of the International Conference in Data Mining*, 2006.

[8] R. P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System," *ACM Transactions on Information Systems*, Vol.27, No.2, Article 12, February 2009.

[9] S. Ahn and S. B. Cho, "Stock Prediction Using News Text Mining and Time Series Analysis," *Proceedings of the KIISE 2010 conference*, Vol.37, No.1, 2010.6, pp.364-369.

[10] V. Sehgal and C. Song, "SOPS: Stock Prediction using Web Sentiment," *Seventh IEEE International Conference on Data Mining – Workshops*, 2009, pp.21-26.

[11] W. Paik, M. H. Kyoung, K. S. Min, H. R. Oh, C. Lim and M. S. Shin, "Multi-stage News Classification System for Predicting Stock Price Changes," *Journal of the Korea Society for Information Managemen*t, Vol.24 No.2, 2007, pp.123-141.

[12] Yu, Y., Y. Kim, N. Kim, S.R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol.19, No.1, 2013.3, pp.92-110.

[13] Yune, H.,H. Kim, J.Y. Chang, "An Efficient Search Method of Product Review using Opinion Mining Techniques," *Journal of KIISE: Computing Practices and Letters*, Vol.16, No.2, 2010.2, pp.222-226.