Stock Price Prediction Through the Sentimental Analysis of News Articles

Jaeyoon Kim, Jangwon Seo, Minhyeok Lee, Junhee Seok

School of Electrical Engineering
Korea University
Seoul, South Korea
jyoonkim@korea.ac.kr
jwein307@korea.ac.kr
suam6409@korea.ac.kr
jseok14@korea.ac.kr

***Abstract*—As people's interest in forecasting stock prices has been recently increased, research on stock price analysis through big data and artificial intelligence has been actively conducted. In this paper, we performed sentimental analysis by building and analyzing a sentimental dictionary with news articles. Through the sentimental dictionary, we can obtain the positive index of news articles for each date. By analyzing the correlation value between the positive index value and the stock return value, we can confirm the utility and possibility of the sentimental analysis in stock market.**

Keywords—Stock price analysis; Sentimental analysis; Sentimental dictionary; Positive index value; Stock return value

I. Introduction

Nowadays, machine learning techniques are being studied in many fields. For stock investors, an accurate prediction of stock price movement may yield profits. However, due to the complexity of stock market data, development of efficient models for predicting is very difficult. In addition, we can choose a variety of data for the research. That's because there are so many factors that contribute to stock price rises and falls. Many machine learning models are being studied such as a naïve Bayesian text classifier[1], Artificial Neural Networks, Support Vector Machines(SVM)[2][3], Random Forest[4], etc. There are many studies that compare the performance of these models.

There is also a sentimental analysis method using Natural Language Processing (NLP), without using various algorithm of such machine learning. The application of this method is such as build models for classifying "tweets" into positive, negative and neutral sentiments [5], examine opinions about the product made in blog posts, comments and reviews [6]. As such, sentimental analysis through Natural Language Processing (NLP) can be very useful in many ways in daily life. In the stock market, text data such as news articles would have a significant impact on stock prices. In this paper, we can confirm whether the data affected the stock price through sentimental analysis.

There are several methods in the sentimental analysis, one of which is Word2Vec [7]. In addition to this method, constructing a sentimental dictionary is another way to do sentimental analysis. There are two approaches. The first is to use KOSAC(Korean Sentiment Analysis Corpus) sentimental dictionary and the second is to construct a new dictionary by using news articles [8]. In this paper, we construct a new dictionary. That is because the words that are mainly used in stock news articles would be different from the words in the KOSAC dictionary. In this point, there is a great need to construct sentimental dictionary which composes of words more related to stocks.

The data needed for the study were downloaded from a website called 'bigkinds'(https://www.bigkinds.or.kr/). Among the various stocks, we gathered news articles from bio-related companies. That is because we thought bio-related stocks would be quite affected by news articles. The sentimental dictionary was constructed with two years of news articles. The test data are news articles about bio related stocks in recent 3 months. Since our data is text data in Korean, we utilize a Korean morpheme analyzer which named KKMA (Kind Korean Morpheme Analyzer).

II. Method

A. Construction of sentiment dictionary

There are three major methodologies for producing sentiment dictionary. Word tagging by hand, methodology using positive index and deep learning such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [9]. In this paper, we suggest the second method that is relatively uncomplicated and intuitively understandable.

The data required for the study are newspaper articles and stock prices for each date. For the stock prices, we use next day

This research was supported by grants from NRF of Korea (NRF-2017R1C1B2002850) and KHIDI (HI19C0360, HI19C0201) as well as a grant from Mirae Asset Global Investments. The correspondence should be addressed to jseok14@korea.ac.kr

978-1-7281-1340-1/19/$31.00 ©2019 IEEE

stock price (NSP) and stock price return (SPR). NSP is an indicator of whether the stock prices have risen the following day. SPR represents the return value between the closing price of the next day and the closing price of today.

Sorting news articles in chronological order before constructing sentiment dictionary. Then, remove the 'stopword' from the content of a news article by hardcoding. This is one of the steps that needs to be taken to prevent unnecessary words from being included in the dictionary. After that, extract words from the contents of a news articles by using a Korean morphological analyzer called KKMA that makes words noun.

To build a sentiment dictionary, we need to know the frequency, positive and positive index (PI) of the words. Frequency means the total number of times the word appears in a news articles. The positive value means the total number of times the word has appeared in a news articles on the date the next day's stock price rises. Finally, the positive index value is obtained by dividing the positive value by frequency. The positive index has a value between 0 and 1. The closer the word's positive index is to 1, the more positive it means.

$$\text{word}(i,\text{j}) = \text{The number of times a word appeared in a news article}$$

$$\text{frequency}(i) \;=\; \sum_{j=1}^{n} word(i,j)$$

$$\text{NSP}(j) \;=\; \begin{cases} \text{1: If the next day's stock price rises} \\ \quad \text{after the article j is posted} \\ \text{0 : next days'stock price decreases} \\ \quad \text{or same after the article is posted} \end{cases}$$

$$\text{Positive}(i) \;= \sum_{j=1}^{n} \{word(i,j) \;\times\; NSP(j)\}$$

$$\text{PI}(i) \;= \frac{\sum_{j=1}^{n}\{word(i,j) \times NSP(j)\}}{\sum_{j=1}^{n} word(i,j)}$$

Some words extracted from the news article, words which have very low frequency can have a high positive value. We excluded these words from the dictionary because it can be a factor that makes it impossible to analyze the fluctuation of the stock price properly.

Finally, the process of making a sentiment dictionary is over. We applied this dictionary to new news articles. First, analyze the morphology of the data to be tested and extract the nouns. Compute the positive index of the text by comparing the extracted nouns with the words of the sentiment dictionary. The positive index of text (PT) is the average value of the positive index of the nouns extracted from the text.

After calculating the positive index of text (PT), calculate the positive index for each date. Daily positive index for each date (DP) is the average of the positive index of news articles published on each date.

$$\text{match}(i, j\,) \;=\; \begin{cases} \text{1 : If the noun j contained in text i} \\ \quad \text{exists in the dictionary} \\ \text{0: In other cases} \end{cases}$$

$$\text{PT}(i) \;=\; \frac{\sum_{j=1}^{n}\{match(i,j) \times PI(j)\}}{\sum_{j=1}^{n} match(i,j)}$$

$$\text{DP}(i) \;=\; \frac{\sum_{j=1}^{n} PT(j)}{n} \qquad \textit{n= number of text in i}$$

B. Dataset

New articles were downloaded from a website called 'bigkinds' (https://www.bigkinds.or.kr/). The news articles are about bio-related companies. The companies are 'Biomed', 'Celltrion', 'Medytox', 'Samsungbiologics' and 'Sillajen'. The dictionary was constructed with two years articles from 2016 March 8 to 2018 December 31. The total number of news articles is 67,968. Those are the train data set. For the test data set, we used news articles for the latest three months. From 2019 January 1 to 2019 March 7. The total number of news articles is 5,047. We used the test data when we analyzed correlation value between the positive index and the stock return value.

III. Experiment Results

Sentiment dictionary consists of words and positive index value (PI). The words were sorted in descending order based on the frequency. Therefore, through the dictionary, we can see which words come out frequently in news articles. The table 1 below is the top 10 words of frequency among the dictionary. Common sense can also confirm that those are the words that will appear frequently when we read news articles about stocks.

TABLE I. Top 10 words in the dictionary

Word	Positive index
일	0.4470
코스	0.4506
피	0.4488
지수	0.4479
시	0.4569
포인트	0.4470
바이오	0.4410
코스피	0.4515
코스닥	0.4410
원	0.4510

If there are too many words in the dictionary, it may include even the words with very low frequency. It can make it impossible to measure the exact positive index. Therefore, in this paper, several experiments were conducted. Changing the number of words in the dictionary. The results are shown in the graph below.

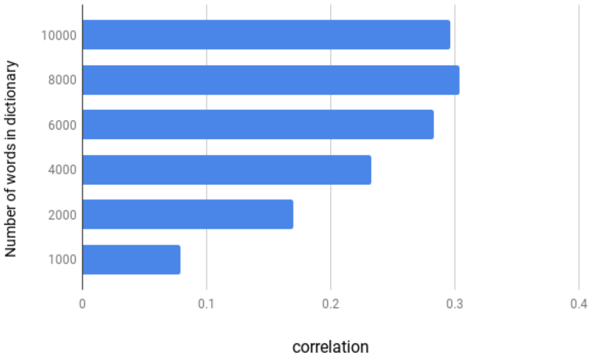

Fig. 1. The value of the correlation here is correlation value between the positive index(PI) and the next day stock return value(SPR)

Fig. 1 shows that the correlation value has increased slightly when there are 8000words compared to 10000 words in the dictionary. However, we can confirm that the correlation value has decreased every time when the number of words decreases. From this result, we can see that it is difficult to accurately predict the stock price fluctuations due to some of the words which have too low frequency. Those words can have high positive value because of the low frequency. In addition, if there are too few words in the dictionary, the probability of finding the word used in the news article in the dictionary getting lower. Then, the correlation value also getting lower. As a result, the experiment found that the 8000 words in the dictionary is the best case.

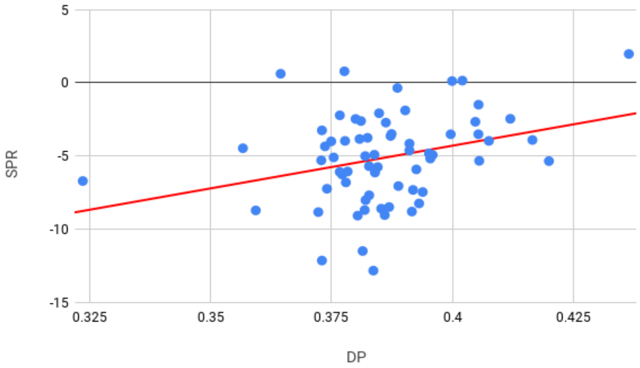

Fig. 2. x-axis is DP and y-axis is SPR. The correlation value is around 0.3034

Fig 2 shows a graph when there are 8,000 words in the sentiment dictionary which is the highest correlation value. The x-axis represents the daily positive index for each date (DP) and the y-axis represents the return value of the next day's stock price (SPR). The correlation value is around 0.3034. This can also be seen through the red trend line.

IV. Conclusion

In this paper, we predict stock price through emotional analysis by using news articles. The method that we used for the emotional analysis is to construct a new sentiment dictionary based on the news articles. Through this, we obtained the correlation between the positive index for each date (DP) and the return value of the next day's stock price (SPR). From this we can see the trend of stock price fluctuations. Based on the experiment, approximately 0.3034 correlation value was checked. Due to the nature of Korean language, it is relatively difficult to derive its high accuracy in using natural language processing methods, rather than English. Therefore, this degree of correlation in the highly complex stock data is thought to be quite meaningful. We confirm that if we remove more unnecessary words through the code and conduct diverse experiments with variety stock items, we can create more efficient sentiment dictionary.

References

[1] G. Gidófalvi and C. Elkan, "Using News Articles to Predict Stock Price Movements", University of California, San Diego, Jun 15, 2001

[2] Y. Kara, M. A. Boyacioglu and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", Published in Expert Systems with Applications, vol. 38(5), May1, 2011

[3] K. Kim, "Financial time series forecasting using support vector machines", Published in Neurocomputing 2003, DOI: 10.1016/S0925-2312(03)00372-2

[4] J. Patel, S. Shah, P. Thakkar and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques", Expert Systems with Applications: An International Journal, vol. 42, no. 1, January 2015, pp. 259-268

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data", LSM '11 Proceedings of the Workshop on Languages in Social Media pp. 30-38

[6] G.Vinodhini and R.M.Chandrasekaran, " Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 6, June 2012

[7] Z. Su, H. Xu, D. Zhang and Y. Xu, "Chinese sentiment classification using a neural network tool – Word2Vec", Expert Systems With Applications, vol. 42, no. 4, pp. 1857-1863

[8] D. Kim, J. Park and J. Choi, "A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles", Korea IT Service Association, vol.12, no. 3 [2014], pp. 211-233

[9] M. Cliché, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", Proceedings of SemEval-2017, 20 April 2017