Extraction of News Articles Related to Stock Price Fluctuation Using Sentiment Expression

Kazuto Tanaka
Major in Computer and Information Sciences
Graduate School of Science and Engineering,
Ibaraki University
email: 20nm715g@vc.ibaraki.ac.jp
4-12-1, Nakanarusawa, Hitachi, Ibaraki, Japan

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
email: minoru.sasaki.01@vc.ibaraki.ac.jp
4-12-1, Nakanarusawa, Hitachi, Ibaraki, Japan

Abstract—**In recent years, various economic reports are published and they are important tools in all markets. However, it is necessary to consult a lot of news articles in order to write an economic report. In this paper, we propose a method for effectively extracting news articles including important events related to the fluctuation of the Nikkei Stock Average by using Japanese sentiment lexicons. The results of the experiments show that the proposed method reduces the number of articles by about half and retrieves relevant documents better than the method using only stock price fluctuations. Therefore, the Japanese sentiment lexicons is effective for extracting news articles including important events.**

Keywords- economic articles extraction; sentiment analysis; Nikkei Stock Average

I. INTRODUCTION

In today's information society, many individual investors try to stay up to date with the latest situations by reading news articles on the web. The news articles describe happenings that are currently occurring in the world and contain information valuable for various purposes. Thus, it is important for investors and traders to be aware of current market situations. However, with the exponentially increasing amount of available news articles, it has become a critical matter to utilize the information from these data in decision making process.

There are various approaches to perform event detection in news articles to help analysts to analyse large amounts of financial news articles. For example, Nakayama et al. propose a method that adds the words appearing in the article and the words appearing in the web page of the company, and uses these words as input for Support Vector Machine (SVM) [1]. In addition, Sakai et al. proposed a method to extract causal information as a causal expression using statistical information and initial cue expressions [2] , and Milea et al. proposed a method using fuzzy grammars [3]. In this existing approaches, we focus on research that analyses the relationship between stock price fluctuations and news articles using text mining techniques. Valuable and significant information related to the financial markets is now available easily on the Internet. Hence, it is neither insignificant nor simple to obtain valuable information and analyse the relationship between the information obtained and the financial markets. However, some existing studies have automatically generated reports by extracting important events and performance factors from previously extracted important news articles. Therefore, there were few studies that automatically extracted important news articles from a large number of news articles.

In this paper, we propose a method to extract important articles related to the fluctuation of the Nikkei Stock Average by using Japanese sentiment lexicons from financial news articles. In the proposed method, we show that the use of the Japanese sentiment lexicons is effective for extracting useful news articles. By using the proposed method, it is possible to automatically extract useful articles for making reports from a large amount of news article data.

The rest of this paper is organized as follows. Section 2 is devoted to the related works in the literature. Section 3 describes the proposed important news extraction method. We describe an outline of experiments in Section 4 and experimental results in Section 5. Finally, we discuss the results in Section 6 and concludes the paper in Section 7.

II. RELATED WORKS AND METHODS

This section presents the related works and methods related to our research.

A. *Related Works*

In recent years, there has been increasing interest in applying neural networks and machine learning techniques to solve this task. First, we introduce some researches related to the analysis of fluctuation of the Nikkei Stock Average.

Nakayama et al. used the Nihon Keizai Shimbun to extract articles that describe events that affect the stock prices of some selected companies [1]. In this paper, they propose a method that adds the words appearing in the article and the words appearing in the web page of the company, and uses these words as input for SVM. This method improved the accuracy by 27.2% compared to the method before adding the words in the web page. However, this experiment does not take into account the combination with the Nikkei Stock Average.

Sakai et al proposed to use Japanese financial news to extract information on the causes of fluctuations in corporate performance [2]. In this paper, causal information is extracted as causal expressions using statistical information and initial cue expressions. As a result, the accuracy of the method was 79.2%, which is better than the result of the conventional method.

Milea et al, proposed to predict the movement of the MSCI Euro Index from the report prepared by the European Central Bank (ECB) [3]. In this paper, they used a fuzzy grammar to create their model. As a result, the results are the same as in previous studies, but the model is simplified.

As mentioned above, there have been many studies on predicting and extracting the causes of fluctuations from various texts, but few studies have focused on extraction of important articles related to the stock price fluctuation to generate investor reports. In this study, we focused on the extraction of important articles related to the Nikkei Stock Average.

Next, we introduce some researches related to the financial text analysis based on sentiment dictionary.

Sato et al. investigated the correlation with stock price fluctuations using a sentiment analysis for news articles in Japan [4]. The results showed that all correlations were low and no correlation could be found between them.

Yazdani et al. proposed a classification method to classify financial news articles into positive or negative class [5]. In their experiments, they used the N-gram models unigram, bigram, and a combination of unigram and bigram as feature extraction, and used Document Frequency (DF) to evaluate the N-gram models and traditional feature weighting methods. The results show that the feature selection and feature weighting methods play a significant role in the negative-positive classification of articles.

Yadav et al. proposed an unsupervised learning model for sentiment classification of financial news articles [6]. The experiment was conducted by using POS based feature extraction with seed set and noun-verb combination in the traditional method. The results showed that using the noun-verb combination was better than the traditional method.

With reference to the above studies, we decided to focus on sentiment expressions.

B. *Sentiment Expression*

There are two types of words: those that give a good impression (positive) and those that give a bad impression (negative). These words are expressed by converting them into numerical values, which is called sentiment information. One of the ways to convert sentiment information is to use a sentiment dictionary. A sentiment dictionary is a dictionary of values determined for each word. The sentiment dictionary we used this time was provided by the Okumura-Takamura Laboratory at Tokyo Institute of Technology [7]. The sentiment dictionary is expressed using numbers between 1 and -1, where the closer the number is to 1, the more positive the word is, and the closer the number is to -1, the more negative the word is. In this study, we use this method to convert news articles into numerical values.

C. *Morphological Analysis*

Unlike other languages, Japanese is characterized by the fact that words are not separated, but are written consecutively. Therefore, it is necessary to divide a sentence into separate words. Morphological analysis is a technology that allows a machine to do this automatically.

In this study, we used a morphological analysis package called janome. This janome uses dictionaries used by a morphological analysis engine called MeCab, which can perform morphological analysis with high accuracy, so it can perform morphological analysis with the same high accuracy as MeCab.

In addition, these morphological analysis engines can not only segment each word, but also analyse the parts of speech and conjugations of the segmented words.

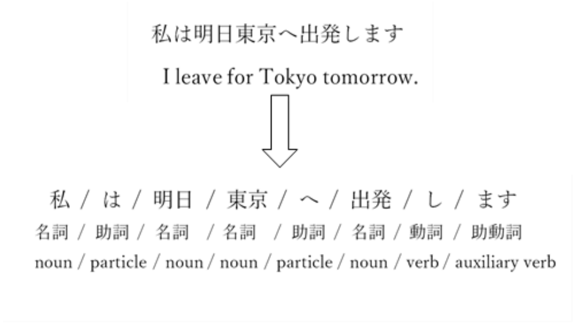

私は明日東京へ出発します
I leave for Tokyo tomorrow.
私 / は / 明日 / 東京 / へ / 出発 / し / ます
名詞 / 助詞 / 名詞 / 名詞 / 助詞 / 名詞 / 動詞 / 助動詞
noun / particle / noun / noun / particle / noun / verb / auxiliary verb

Figure 1. Examples of Morphological Analysis

As an example, the morphological analysis of "私は明日東京へ出発します"(I am leaving for Tokyo tomorrow) is shown in Figure 1.

III. Extraction Method Using Sentiment Expression

In this section, we present a method to extract news articles that have information related to the fluctuation of the Nikkei Stock Average.

A. *Overview of the Proposed Method*

Figure 2 shows a rough sequence of the proposed method.

"Headline" in the figure is the title of the news article. "Date" is the date when the news article was published. "Open" is the opening price of the Nikkei Stock Average. "Close" is the closing price of the Nikkei Stock Average. "Positive", "neutral", and "negative" are numerical values for each word in the headline. "Positive score", "neutral score", and "negative score" are each summarized for each day.

First, we select data from article data and stock price data. Since the article data contains some articles that may be noisy, we select articles under certain conditions. The details are introduced in Section B. We considered the correct article to be an article about a major event related to the fluctuation of the Nikkei Stock Average.

Next, the selected articles are polarity transformed. Before polarity transformation, it is necessary to divide the keywords of the articles into the smallest units of words. Therefore, the keywords are first analysed by morphological analysis, and then converted into numerical values by the polarity dictionary. The detailed conversion method will be explained in Section C.

Then, we use the obtained polarity to calculate whether the article is a positive or negative article on stock prices. The method is to calculate the three percentages of whether the article is positive, negative, or irrelevant (neutral). The detailed calculation method will be explained in Section D.

Finally, we extract articles from the results of polarity representation. Articles with positive values in the top 75% and negative values in the bottom 25% are extracted. At this time, if duplicates often occur in both values, the duplicates are removed.

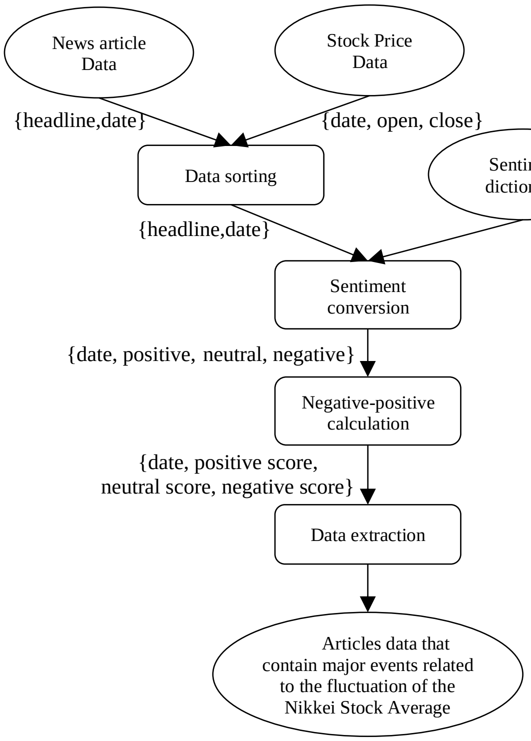

News article Data
Stock Price Data
{headline,date}
{date, open, close}
Data sorting
Sentiment dictionary
{headline,date}
Sentiment conversion
{date, positive, neutral, negative}
Negative-positive calculation
{date, positive score, neutral score, negative score}
Data extraction
Articles data that contain major events related to the fluctuation of the Nikkei Stock Average

Figure 2. Flow of the Proposed Method

We compare the percentage of articles that contain major events related to the fluctuation of the Nikkei Stock Average between the extracted set of articles and the unextracted set of articles to verify the effectiveness of sentiment expression.

B. Preprocessing

First of all, we need to sort out the data from Nikkei QUICK News, which we use, because there are some news articles that are not related to the Nikkei Stock Average, such as sports articles. The first step is to limit the news articles to two sources, "QUICK" and "NQN". Since these two sources only provide news about stocks, we limited ourselves to these two sources.

Next, news articles distributed by "QUICK" and "NQN" include a mixture of news articles summarizing the past and news articles about a company's sales. These articles can have a negative impact on the use of sentiment expressions. Therefore, we removed such articles.

In addition, we considered that news articles related to the fluctuation of the Nikkei Stock Average were likely to be transmitted on the days when major fluctuations occurred. Therefore, we limited ourselves to the news articles that were published on days when the Nikkei Stock Average was more than 1% or less than -1% compared to the previous day. In this paper, such days are referred to as Fluctuation Day.

Finally, in some of the data used in this study, the keywords of the news articles are empty. In this paper, the articles with empty keywords are also removed in order to use these keywords for polarity conversion.

C. Sentiment Conversion

Sentiment conversion of articles limited by preprocessing into three categories: positive, negative, and neutral. Neutral means a word that is neither positive nor negative. For this conversion, we morphologically analyse the keywords in the news article. This is because the words present in the sentiment dictionary are the smallest words registered in natural language. The words obtained from this morphological analysis are compared with the sentiment dictionary and converted into numerical values. Positive values are considered positive, and negative values are considered negative. If the word does not exist in the sentiment dictionary, it is considered a neutral word and converted to 0. If the word is not in the sentiment dictionary, it is considered a neutral word and converted to 0. Furthermore, words that are replaced with 0 when converted to numerical values are treated as neutral words. This is done for all articles.

D. Negative-Positive Calculation

Each of the positive, negative, and neutral values converted in the previous section are combined into one. In this experiment, we averaged the positive, negative, and neutral values for each article and used the individual values to find the percentage of each. The reason for using percentages is that the number of keywords in each article is different, and it is easy to extract articles with many keywords using only the average. Therefore, this time we will use the average value to obtain the percentage. Specifically, we used the formula below to calculate the percentage. Since neutral words are less important than positive and negative words, the average of all neutral words is set to 1 if neutral words exist, and 0 otherwise.

- Positivity formula

$$positive\_score = \frac{\overline{positive}}{\overline{positive} + \overline{negative} + \overline{neutral}}$$

- Negatives formula

$$negative\_score = \frac{\overline{negative}}{\overline{positive} + \overline{negative} + \overline{neutral}}$$

- Neutral formula

$$neutral\_score = \frac{\overline{neutral}}{\overline{positive} + \overline{negative} + \overline{neutral}}$$

IV. Experiments

The experimental procedure based on the proposed method is shown.

A. Data Set

In this study, we used data compiled from Nikkei QUICK News for 2016 and 2017 and the prices of the Nikkei Stock Average for 1747 days from 2011 to 2017, as well as a polar dictionary provided by Okumura and Takamura laboratories at Tokyo Institute of Technology [7].

B. Settings

In this experiment, we decided to focus on three important events: the Ise-Shima Summit, the U.S. presidential election, and the North Korean missile launch. Therefore, the experiment was conducted every month in the month in which these events occurred. In doing so, articles that included "サミット"(summit), "選挙"(election), and "地政学リスク"(geopolitical risk) as keywords in their articles were considered as correct articles.

- May 2016: Ise-Shima Summit
- November 2016: U.S. presidential election
- August 2017: geopolitical risk

In addition, the U.S. is often involved in the fluctuations of the Nikkei Stock Average. Therefore, we conducted an experiment with an article set that was limited to articles that contained "米国"(U.S.) or "アメリカ"(America) in the keywords of the articles. In this study, we refer to this set of articles as the set of articles containing "U.S.".

Thus, this study was conducted using two patterns of data for each of the three events.

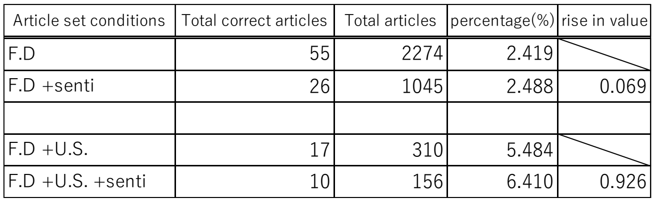
V. Results

The results of an experiment conducted based on the experimental method are shown below.

A. Results of the May 2016 Article Set

The results done in May 2016 are shown in Table 1 below. Also, there are six Fluctuation Days: 2, 10, 13, 17, 25, and 30.

TABLE I. RESULTS OF THE MAY 2016 ARTICLE SET



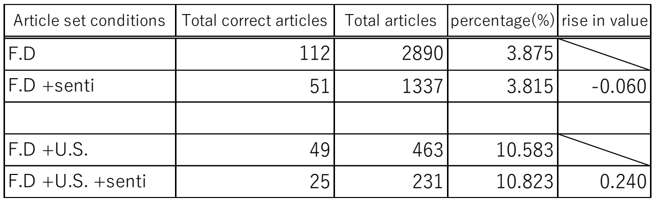
Article set conditions	Total correct articles	Total articles	percentage(%)	rise in value
F.D	55	2274	2.419	
F.D +senti	26	1045	2.488	0.069
F.D +U.S.	17	310	5.484	
F.D +U.S. +senti	10	156	6.410	0.926

In the table, "F.D" refers to Fuctuation Day. Also, "F.D + U.S." refers to the set of articles that include "U.S.". In addition, "F.D +senti" and "F.D +U.S. +senti" are the results of extraction using sentiment expressions. "Total correct articles" is the number of correct articles for each condition. "Total articles" is the number of articles when the data is sorted under each condition. "Percentage" is the ratio of correct articles to the total number of articles. "Rise in value" is the increase in percentage after the experiment.

B. Results of the November 2016 Article Set

The results done in November 2016 are shown in Table 2 below.

TABLE II. RESULTS OF THE NOVEMBER 2016 ARTICLE SET



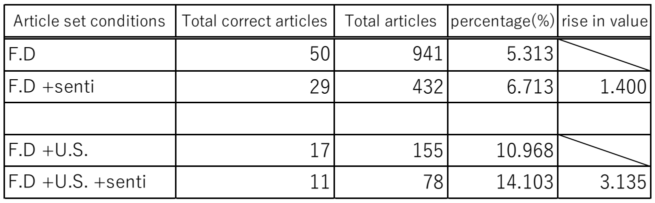
Article set conditions	Total correct articles	Total articles	percentage(%)	rise in value
F.D	112	2890	3.875	
F.D +senti	51	1337	3.815	-0.060
F.D +U.S.	49	463	10.583	
F.D +U.S. +senti	25	231	10.823	0.240

It is the same as Table 1 with respect to the making of the table. In addition, there are seven Fluctuation Days: 2, 4, 7, 9, 10, 14, and 16.

C. Results of the August 2017 Article Set

The results done in August 2017 are shown in Table 3 below.

TABLE III. RESULTS OF THE AUGUST 2017 ARTICLE SET



Article set conditions	Total correct articles	Total articles	percentage(%)	rise in value
F.D	50	941	5.313	
F.D +senti	29	432	6.713	1.400
F.D +U.S.	17	155	10.968	
F.D +U.S. +senti	11	78	14.103	3.135

It is the same as Table 1 with respect to the making of the table. In addition, there are six Fluctuation Days: 9, 15, and 18.

VI. Discussions

In this section, we discuss the experimental results and show the effectiveness and problems of polarity representation.

A. Validity of Sentiment Expression

The results of Section 5 show that all article sets have a better percentage of articles containing information related to the fluctuation of the Nikkei Stock Average, except for the extraction performed on the article set for the Fluctuation Day of November 2016. In addition, this extraction method reduced the original article set by almost half, which means that we were able to reduce the number of articles that need to be manually examined.

The reason for the worse results in November 2016 was that there were articles on Clinton's email problem and articles on the current status of the presidential election that did not include "選挙"(election) as a keyword. Therefore, it is thought that such articles were extracted rather than articles containing "選挙"(election) and thus the good results were not obtained.

Considering the above, we believe that the use of sentiment expressions is effective in extracting articles that have information related to the fluctuation of the Nikkei Stock Average.

B. *Inadequate Preprocessing*

From the results of Section 5, we can see that both the percentage and the upward value of the set of articles including "U.S." are better than the set of Fluctuation Day. This was probably due to the fact that we were able to remove articles that were not so relevant to the fluctuation of the Nikkei Stock Average, such as news about companies, since the percentage was better just by limiting the articles to those that included "U.S.". Therefore, the pre-processing of this experiment alone is not enough to sort out the data.

In the future, it is necessary to review the data selection process and to improve the weighting of each article.

C. *Optimal Threshold for Front-to-back Comparison*

This study was limited to the days when the Nikkei Stock Average was more than 1% or less than -1% compared to the previous day. Therefore, differences occur in the number of days examined in different months. In this study, there were six days in May 2016 and seven days in November 2016, but in August 2017, there were only three days, less than half. Furthermore, in August 2017, only the 9th, 15th, and 18th were Fluctuation Day, so it was not possible to take articles from the beginning and end of the month. Therefore, it is necessary to adjust this threshold of the Nikkei Stock Average compared to the previous day from month to month.

VII. Conclusion

In this study, news articles related to the fluctuation of the Nikkei Stock Average were extracted from the price information of the Nikkei Stock Average and Nikkei QUICK News by using sentiment expressions. The news articles were converted into numerical values from the sentiment dictionary, and the articles were extracted from these values. To verify the effectiveness of the extraction method using sentiment expressions, we compared the percentage of news articles related to the fluctuation with the original set of articles. As a result, the number of articles in the article set was reduced by almost half when the extraction method using sentiment expressions was used, and the percentage of articles included in the articles related to fluctuations also increased. Thus, we were able to confirm the effectiveness of polar expressions.

Future tasks include improving the pre-processing data selection method, weighting of each article, and adjustment of the fluctuation date threshold for each month.

References

[1] M. Nakayama, H. Sakaji, and K. Katsuta, "Hiroyuki Sakai: Extraction of Important Articles that Influence the Stock Price of Companies from Financial Articles," The Journal of the Faculty of Science and Technology, Seikei University, Vol.51, No.2, pp.53-60, 2014.

[2] H. Sakai, and S. Masuyama, "Cause Information Extraction from Financial Articles Concerning Business Performance," IEICE Transactions on Information and Systems, Vol.E91-D, No. 4, pp. 959–968, 2008

[3] V. Milea, N. M. Sharef, R. J. Almeida, U. Kaymak, and F. Frasincar, "Prediction of the MSCI EURO index based on fuzzy gram- mar fragments extracted from European Central Bank statements," International Conference of Soft Computing and Pattern Recognition, pp.231–236, 2010.

[4] K. Sato, T. Odaka, J. Kuroiwa, and H. Shirai, "Study on the Correlation between Stock Price and Web Data Using Negative-Positive Analysis," Memoirs of the Graduate School of Engineering, University of Fukui, Vol. 63, pp.75-86, 2015.

[5] S. F. Yazdani, M. A. A. Murad, and N. M. Sharef, "Sentiment Classification of Financial News Using Statistical Features," International Journal of Pattern Recognition and Arti¯cial Intelligence Vol. 31, No. 3, 2017.

[6] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, "Sentiment analysis of financial news using unsupervised approach," Procedia Computer Science Volume 167, pp.589-598, 2020.

[7] H. Takamura, T. Inui, and M. Okumura, "Extracting Semantic Orientations Using Spin Model," Journal of Information Processing Society of Japan, Vol.47 No.02, pp. 627-637, 2006.

ISBN: 978-1-61208-850-1