Detection Model for Correlation between SNS Spikes and Stock Price Movement

Naoki Itoh, Yuriko Yano, and Yukari Shirota

***Abstract*—Stock price prediction has become an important research theme in text mining application fields. Many models of sentiment analyses for the prediction have been published. If they found SNS volume spikes on companies or products, many stock investors might quickly sell the stocks even if they cannot predict whether the stock price direction is an increase or a decrease. In the paper, we shall propose a detection model for the clear-cut correlation between a SNS spike and stock price movement. If we find a SNS spike, firstly, topic extraction is conducted on the SNS text data to remove the noise data to extract a purely breaking topic. Then, from the breaking topic distribution, we make the differential equation. Finally, we determine whether the solution data matches the actual stock price data.**

***Index Terms*—Stock prediction, text mining, breaking news, SNS analysis, SNS volume spikes.**

I. Introduction

Stock price prediction has become an important research theme in text mining application fields. Many models of sentiment analyses for the prediction have been published. In order to detect a rapid change on stock prices, it is useful to monitor SNS data such as twitter. If they found SNS volume spikes (or breaking SNS data) on companies or products, many stock investors might quickly sell the stocks even if they cannot predict whether the stock price direction is an increase or a decrease. Although a sentiment analysis by text mining helps them analyze contents of the SNS spikes, it is difficult to predict the stock price direction.

Another problem is to detect a cause and effect relationship between the SNS spikes and the stock price data. Like measurement of price sensitivity, a product or a company has its own SNS spikes sensitivity. If the SNS spike sensitivity is low, a number of messages are issued on a topic, there would be no effects on the stock price movement.

In the paper, we shall propose a detection model for a clear-cut correlation between a SNS spike and stock price movement. In the next section, we shall describe related work. In Section 3, we shall, as a concrete example, show a clear-cut impact on stock prices by SNS spikes of vehement claims. Using the real data, we shall show our measurement method of a SNS spike sensitivity. Finally, we shall conclude the paper.

Manuscript received March 10, 2018; revised June 30, 2018. This work was supported in part by GEM (Gakushuin University Research Institute for Economics and Management) as the GEM project 2018.

The authors are with Faculty of Economics, Gakushuin University, 1-5-1 Mejiro, Toshima-ku, Tokyo 171-8588, Japan (e-mail: 14022027@gakushuin.ac.jp, 16122006@gakushuin.ac.jp, yukari.shirota@gakushuin.ac.jp).

II. Related Work

We shall survey the related work. Stock price prediction We shall suevey the related work. Stock price prediction has become an important theme in text mining application fields. Because it is beneficial to predict stock market future trends, many prediction models based on sentiment analysis of some text data and historical stock market prices have been published [1]-[3]. Because there are a lot of noises in SNS data such as tweets, we know that an analysis of only Tweets data cannot make execellent performance. Therefore, Patric uses stock market news reports as reliable contents [4]. Jageshwer and Shagufta use the impact of financial news [5]. Sadi uses economic news to investigate the correlation between the news and stock price time series data [6].

Hana and Hasan use both news articles and breaking (or bursting) Tweets data to predict whether a stock price direction is an increase or a decrease. The information they use is 50,000 news articles collected from NASDAQ website matching 30 stock components in Dow Jones Index (DJI). The research results showed that if the tweet volume breakout information is added, then the prediction performance becomes better.

The breaking data on Tweets can be found at natural disasters such as earthquakes and accidents/incidents. Daniel *et al.* called them an event detection [7]. The events include large parties, sporting events, exhibition, promotion of products, accidents and political campaigns, in addition to the natural disasters. Daniel *et al.* for example presented company events with the appropriate calculation of scores for the sentiment analysis tools; for example, concerning the event titled “The biggest pixels thing by Apple makes HTC look good. Very good”, the all four tools classified the event to be positive because the contents can be easily interpreted. However, there are tough tweets which the tools cannot easily identified as positive ones or neutral ones [7].

III. Case Study Data

In the section, we shall explain the case study data that we shall use.

In the paper, we use data concerning a game software called THE IDOLM@STER. THE IDOLM@STER is very popular in Japan among male game players. THE IDOLM@STER is a Japanese raising simulation and rhythm video game series created by Bandai Namco Games (formerly Namco). Owing to the popularity of THE IDOLM@STER, the stock price of the Bandai Namco Games had increased. However, on September 18th 2010, the second series of the game titled THE IDOLM@STER (abbreviated as IDOLM@STER2) had be announced which

doi: 10.18178/ijtef.2018.9.6.626

occurred a great bursting data spike on SNS in Japan (See Fig. 1). Consequently, the company stock prices greatly decreased in the September.

Let us explain the situation in which the users monitor the SNS spikes. Fig. 1 shows the number of messages concerning IDOLM@STER2. The SNS data we used in the analysis are data from Yahoo Answers which in Japanese we call "Yahoo chiebukuro" (https://en.wikipedia.org/wiki/Yahoo!_Answers). Yahoo Answers is a community-driven question-and-answer (Q&A) site or a knowledge market from Yahoo! that allows users to both submit questions to be answered and answer questions asked by other users. In Japan, Yahoo Answer is the most popular Q&A website. The Q&A sentences are written in Japanese. We conducted a retrieval by the keyword "IDOLM@STER2".

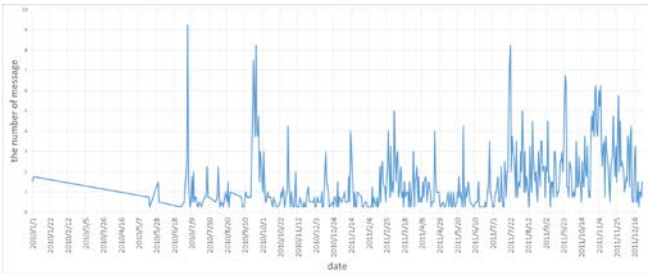

Fig. 1. The number of messages concerning IDOLM@STER2.

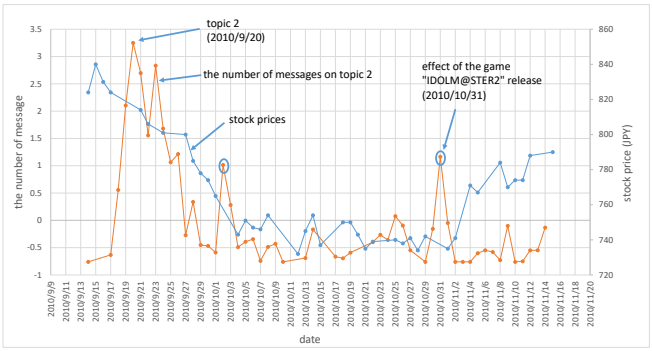

Fig. 2. The number of messages concerning the topic 2 of IDOLM@STER2 and the Bandai stock price movement.

The number of messages on Yahoo Answer is shown in Fig. 1 in which the *x* axis is date and *y* axis is the frequency. As shown in Fig. 1, sometimes spikes are observed. When a data spike is found, we would like to check whether the data spike would have effects on the stock prices or not. If there was a drastic stock price movement, there would be a strong correlation between the data spike and the stock price change. We would like to investigate the followings:

- The correlation is a true one or just coincidences?
- If a real cause and effect relationship exists, what is the key reason or topic?

If the real correlation exists, because SNS messages include a lot of noise data, we will have to remove the noise data from the SNS messages, so that we could obtain the key topic that had effects on the stock price. For example, Fig. 1 shows several data spikes.

However, the data spike on around September 18th 2010 only had drastically effects on the stock prices; other spikes did not make large effects on the stock prices.

In addition, the data spike in September 2010 included many other noises. The noise messages should be deleted by topic extraction, so that we can focus on the target key topic. This kind of monitoring and detection system should be run automatically. There are a large number of SNS spikes anytime on the SNS sites. The human operators cannot monitor the big amount data.

Then, to implement such an automatic monitoring system, what human operators should do? In advance, we should make several search key words for a company. Monitoring the SNS data by using the search key words, if we found a data spike, then the system should check that cause and effect relationship between the data spike and the stock price. To implement the monitoring system, we need a detection model for a clear-cut correlation between a SNS spike and stock price movement.

IV. Correlation Detection Method

In the section, we shall propose and explain the correlation detection method. If we find a SNS spike, firstly topic extraction should be conducted on the SNS text data to remove the noise data to extract a purely breaking topic. Then, from the breaking topic distribution, we make the differential equation. Finally, we determine if the solution data matches the actual stock price data.

A. Topic Extraction

We shall conduct topic extraction to remove the noise data for just one breaking topic. We constructed the topic extraction using the LDA (Latent Dirichlet Allocation) model. The LDA model is a widely-used multi-topic document model based on Baysian inference method [8]. The Markov chain Monte Carlo methods (MCMC) algorithm we used on the LDA model is Gibbs sampling [9]. The Gibbs sampling is widely used. The algorithm is visually explained in [10]. As the program of the LDA with the Gibbs sampling, we used the R packaged offered by "The Comprehensive R Archive Network abbreviated as CRAN titled "lda: Collapsed Gibbs sampling methods for topic models" developed by Jonathan Chang. In LDA model, we must decide in advance the number of topics. We use noun-noun bi-gram terms instead of one term so that the results can be more informative.

The SNS data is one from Yahoo Answers. The original texts are Japanese, and we translated them into English terms there. The period is from January 1st 2010 to February 20th 2011. The total number of messages is 3,350 and the total number of Japanese characters is about 733,000 (no space character is included) which is a very large number, because Japanese characters include informative Chinese characters. We set the number of topics to be five, because the five topic extraction made the clearest result. Among topic 1 to topic 5, we decided the topic 2 must be the target topic.

In Fig. 2, the topic 2 message number movement and the stock price movement are illustrated. There the number of messages on topic 2 is standardized during 14th Sept. 2010 to 14th Nov. 2010; on the data during the period the average value is set to be zero and the standard deviation is set to be one.

We found a spike around on September 20th 2010. Before we conducted the data analysis, the game fans had already insisted on the strong relationship between the announcement of IDOLM@STER2 and the stock price downfall. However,

we need a convincing proof on that to say so. Therefore, we shall propose a detection model to prove the clear-cut correlation between a SNS spike and stock price movement.

The topic 2 includes the following terms: “Ryugu Komachi” (the ex-idol name), “iDOLM@STER”, “male idol”, and “impossible to produce”. In the games, it is called “to produce an idol” to raise an idol. Concerning the incident, we interviewed the game fans and we retrieved web data on that because we had no knowledge about the game and the incident. Then we found the following things. The game company then announced that the plot of THE IDOLM@STER2 and the game plot had been largely changed from the first series. There some idols became ex-idols, which made the enthusiasts of the idols get very angry. Other reason of their wrath was that male idols were newly added in the second series, which also made the fans get worried and cross. Let us see the extracted terms in the topic 2.

In Fig. 3 to Fig. 5, the nearest terms of the keywords “Ryugu Komachi”, “make idol”, and “impossible to produce” are shown. To obtain the nearest terms, we used the word2vec for the analysis [11]. Word2vec, based on the two-layer neural network, makes the word vectors which correspond to principal components so that they can represent words in the vector space. In advance, we make a vector representation model from the input text file. The output is the vector representation model (vector space) in which each word is represented as a vector.

Using the vectors, we can calculate the similarity level between two words. Then, given a word, we can extract the similar word list to the word. Then we can get the subset of the vector space of the extracted word list. For the subset vector space, we shall conduct visualization using t-SNE (t-Distributed Stochastic Neighbor Embedding) as shown in Fig. 3 to Fig. 5 The t-SNE conducts dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets[12]. In the paper, the number of dimensions are reduced from 100 to two. The two axis are the two principal component directions that t-SNE extracted. These kinds of visualization enable us to easily find the related terms of the keyword.

There are some key extracted words illustrating this game’s feature. The most frequent word, “Ryugu Komach”, is one of the most popular idol groups in the game made up of three girls. And this word is related to “impossible to produce.” THE IDOLM@STER2 do not allow producers to produce this idol group, though they were very popular characters (in this game, a “producer” means a game player). On the other hand, the third most frequent word, “male idol,” is corresponding to that a male idol is newly added in this series, which producers did not want.

THE IDOLM@STER had been popular for more than ten years, and its fans are almost male. We found that the angry fans and supporters must have written many vehement antagonism messages against the second series. Consequently, the number of messages concerning THE IDOLM@STER2 had been rapidly soring. They even tried to make a boycott campaign in social media.

In addition, we found other two spikes on October 2nd and 31st (See Fig. 2). The first spike may have no effects on the stock price movement. However, the latter spike made the stock price increased again. We can surely see the correlation between the spike and the stock price increase. To survey the acceleration reason, we read the text messages around on October 2 and 31 and we found that a game titled "IDOLM@STER2" for Play Station 3 was released on October 27th. The Play Station 3 was a very popular game machine and the number of users is large. Therefore, the game release must have increased the SNS messages and the stock price.

On the other hand, around on October 2nd, the game release was announced and the game contents were talked a lot on SNS. The number of messages had increased but the effects were not so large; the stock price had not been increased again but the decrease then had stopped.

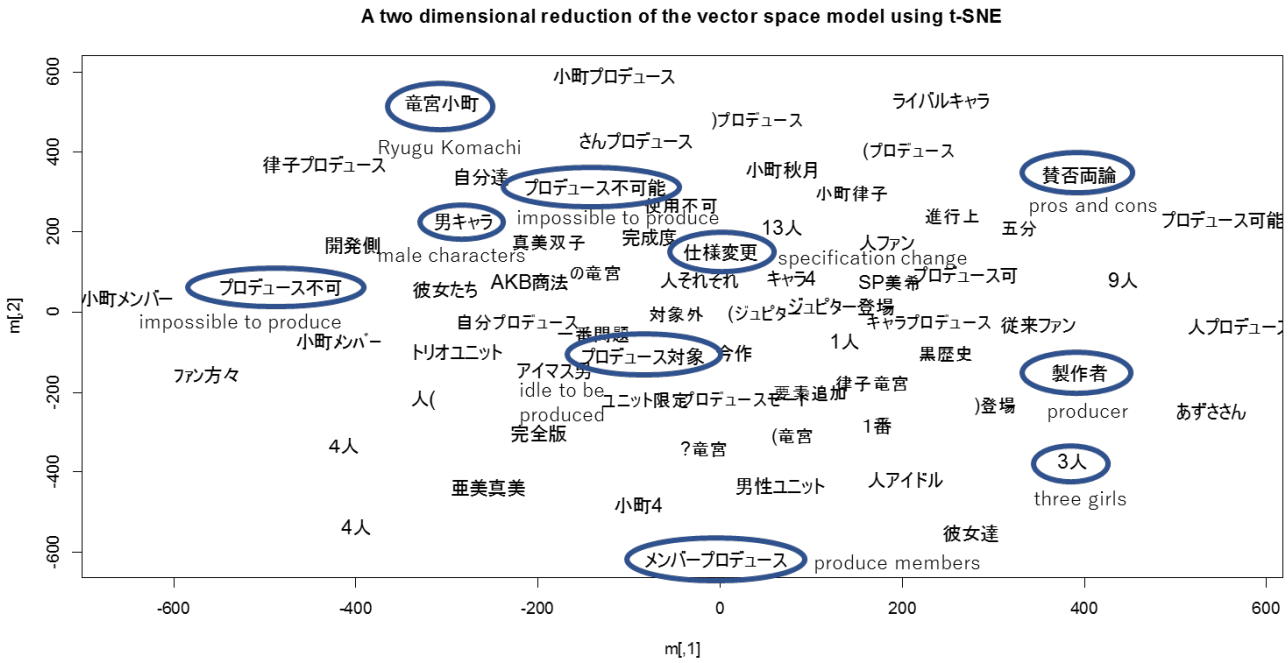

Fig. 3. The nearest terms of “Ryugu Komachi” (the ex-idol name).

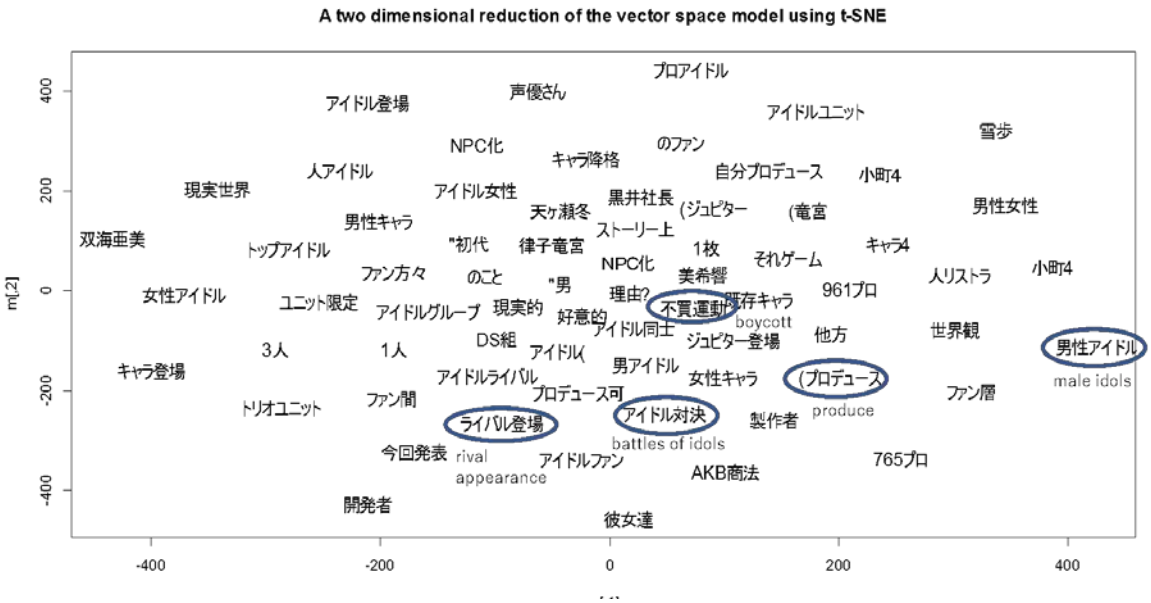

Fig. 4. The nearest terms of “male idols”.

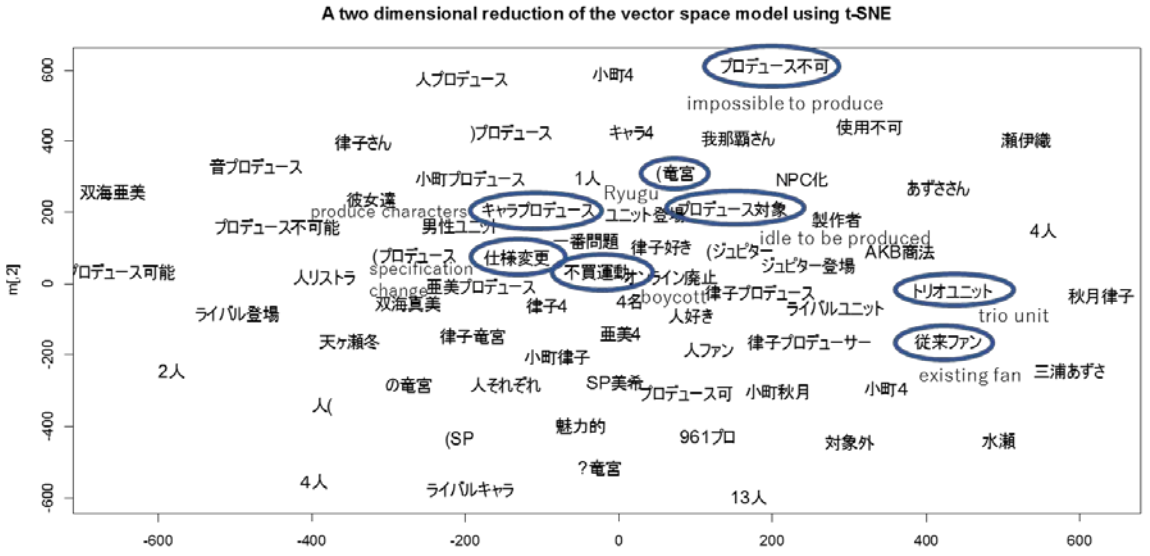

Fig. 5. The nearest terms of "impossible to produce".

B. Differential Equation

We suppose that the SNS volume is a downward acceleration on the stock price movement. In this case, we suppose the differentiation equation during the spike period as follows where S represents the stock price, t is time, and k is a constant number:

$$\frac{d^2S}{dt^2} = -k$$

In other words, a step shaped impulse with a fixed amplitude k. The solution function of S becomes a function $S(t) = -\frac{k}{2}t^2 + at + b$ where a and b are constant values. Input the 12 business day data, we found the fitting function of S by the least square method as follows:

$$S(t) = -0.1748t^2 + -5.5804\,t + 842.4090$$

In Fig. 6, the fitting function is shown with the real stock price movement. The constant k is found to be 0.3496.

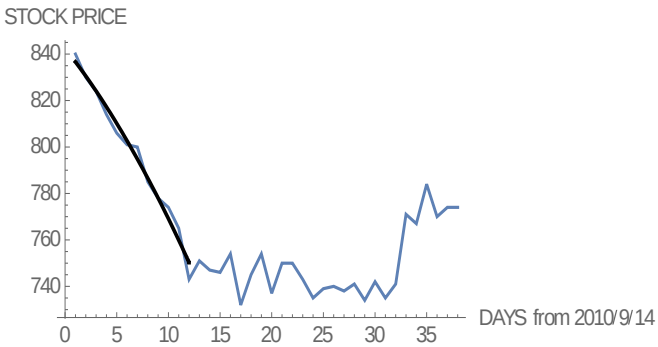

Fig. 6. The fitting function and the actual stock price movement.

C. Evaluation of Similarity Level

Visually we found that the fitting function is very similar to the actual stock price movement during the period. However, we need to conduct the procedure automatically without a human help; a similarity measurement method is

required. The similarity level can be calculated as the *Pearson's correlation coefficient* as far as both data are in advance standardized; the mean value is 0 and the variance is 1 [13]. The result of the Pearson's correlation test which tests whether the vectors v1 and v2 are linearly independent in Table I.

TABLE I: THE PEARSON'S CORRELATION

	Statistic	P-Value
Pearson Correlation	0.990297	6.72444×10^{-10}

The Pearson's correlation coefficient is 0.99. As the range of the Pearson's correlation coefficient is from -1 to 1, the value shows that this fitting function is so similar to the actual stock price data. Since the P-Value is nearly zero, the null hypothesis of independency is rejected. As a result, we can say that the fitting function data is very similar to the real stock price data. Thus, we can say that the strong relationship exists between the spike and the stock price decrease.

V. RESULT ANALYSIS

We proposed the detection method for a clear-cut correlation between a SNS spike and stock price movement. The method is conducted as follows:

1) *Topic extraction:* Using the topic extraction, select the purely breaking topic by removing the other topics as noise data.
2) *Differencial equation:* Suppose that the SNS volume is a downward acceleration on the stock price movement. Set the differencial equation on the stock price S(t) and solve it so that the fitting function can be obtained.
3) *Evaluation of Similarity Level:* Calculate the similarity level between the fitting function and the stock price movement. If the similarity level is high, we can say that there is a strong relationship between the two data. Otherwise if the similarity level is not high, the correlation is neutral.

In general, stock prices of digital contents sales companies are sensitive, we think. Other example of the digital contents was shown in [14]. In [14], an upward acceleration we could find. The virtual marketing tends to create enormous demand. On the other hand, the violent claiming and boycott campaign may decrease the stock prices rapidly. Does the stock holder of such companies hold interests in the sales products? In case of the IDOLM@STER2 incident, we thought that the consumers of the game and the stock investors were exclusive groups. The reason why is there is no term related to stocks or its prices in the SNS messages.

We guessed that the stock holders were not interested in the game and that they must be so sensitive only on SNS spikes, so that just a SNS spike can trigger their selling actions to stop their loss. We tried to investigate the net sales amount of the game at the time using the company's financial report in which the sales amount was reported every quarterly. Then, there was no decrease of the sales amount. From the data, however, we cannot say that there was no impact on the sales amounts of IDOLM@STER by the incident because the company offered other popular games such as Gandum, and the company does not make each game's sales data. In addition, the actual start of the sales of IDOLM@STER2 was in February the next year 2011.

We condacted the same analysis on other products, for example, HEATECH which is a warm underware. In the case of HEATECH, we could see a steady and gradual inceease, compared to the IDOLM@STER2 [15]. There exist several SNS spikes for each winters and then in the winter the large SNS spike appeared.

VI. CONCLUSIONS

Stock price prediction has become an important theme in text mining application fields. Using sentiment analysis techniques, a trading robot or human trader always surveys consumer behaviors over time from buzz marketing sites in order to make decisions on selling or buying the stocks. If the boycott campaign happens on the web, the incident would easily make stockholders sell the stocks as soon as possible to minimize the loss. Currently one of the most effective information to predict stock prices is a SNS breaking spike. The sentiment analysis by text mining helps us identify whether the SNS contents are positive or negative or neutral on the company or its product. In the paper, we propose another point whether the SNS spikes have positive or negative or neutral effects on stock prices. If we detect the strong relationship between them, the stock holders have to do a quick action. For the purpose, the detection method for a clear-cut correlation between a SNS spike and stock price movement is described in the paper.

REFERENCES

[1] L. Bing, K. C. C. Chan, and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company," in *Proc. 2014 ICEBE Conf.*, 2014, pp. 232-239, 2014.

[2] Y. E. Cakra and B. D. Trisedya, "Stock price prediction using linear regression based on sentiment analysis," in *Proc. 2015 ICACSIS Conf.*, 2015, pp. 147-154, 2015.

[3] H. Alostad and H. Davulcu, "Directional prediction of stock prices using breaking news on Twitter," *Web Intelligence (2405-6456)*, vol. 15, no. 1, pp. 1-17, 2017.

[4] P. Uhr, J. Zenkert, and M. Fathi, "Sentiment analysis in financial markets A framework to utilize the human ability of word association for analyzing stock market news reports," in *Proc. 2014 SMC Conf.*, 2014, pp. 912-917, 2014.

[5] J. Shriwas and S. Farzana, "Using Text Mining and Rule Based Technique for Prediction of Stock Market Price," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 1, pp. 245-250, 2014.

[6] E. S. Sadi, C. Mert, K. Al-Naami, N. Ozalp, and U. Ayan, "Time Series Analysis on Stock Market for Text Mining Correlation of Economy News," *International Journal of Social Sciences and Humanity Studies*, vol. 6, no. 1, pp. 23, 2002.

[7] M. Daniel, R. F. Neves, and N. Horta, "Company event popularity for financial markets using Twitter and sentiment analysis," *Expert Systems with Applications*, vol. 71, no. Supplement C, pp. 111-124, 2017.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[9] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[10] Y. Shirota, T. Hashimoto, and B. Chakraborty, "Visual Materials to Teach Gibbs Sampler," *International Journal of Knowledge Engineering*, vol. 2, no. 2 & 3, pp. 92-95, 2016.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781*, 2013.

[12] L. V. D. Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221-3245, 2014.

[13] Y. Okuse and T. Kuboyama, *Practical Data Analysis for* Economics *and Management (in Japanese)*, Kodansha, 2012. 1.

[14] S. Tashiro and Y. Shirota, “Pattern Analysis of Stock Price Increase by Hit Products -- Case Study of LOVELIVE! --,” *IEICE Tech. Rep.*, vol. 118, no. 107, DE2018-1, pp. 1-4, 2018, written in Japanese.
[15] T. Yoshino, A. Takura, and Y. Shirota, "Marketing Awareness by Social Network: A Case Study of HeatTech Products," in *Proc. of The Ninth IEEE International Conference on Awareness Science and Technology (ICAST 2018)*, September 19-21, 2018, pp. 128-132, 2018.

Yuriko Yano is a postgraduate school student at Gakushuin University in Tokyo, Japan. She analyzes effects of natural disasters such as the great east Japan earthquake and Thai flood 2011 on Japanese economy using SVD (Singular Value Decomposition) and text mining technologies as a master's thesis for MBA. Stock prices growth pattern by the emergency demand after the great East-Japan earthquake (Kenji Yamaguchi, Yuriko Yano and Yukari Shirota) won the "Best Poster Award" in ACIIDS 2017. Her fourth research paper, "Differences between Japanese Tipping Customs, Kokorodzuke, and Western Tipping Customs -A Discovery of Japanese Complicated Cultural Features-" won the "Distinguished Paper Award" in International Conference on Business and Information 2017 (BAI 2017). She has dedicated herself to study SVD and text mining to stimulate economics in Hawaii. She is taking an MBA on March 20th in 2018 from Gakushuin University.

Naoki Itoh was graduate from Department of Management, Faculty of Economics Gakushuin University in March 2018. He is interested in economics data analysis by using Mathematica. Especially using both a stock prices movement and a text mining from SNS data on the web, extraction of features on each companies movements is interesting for him. Now he is interested in machine learning on web data analysis.

Yukari Shirota is DSc. Prof. of Faculty of Economics, Gakushuin University. Research fields are visualization of data on the web. data visualization, social media analysis, and visual education methods for business mathematics. For over 19 years, she has developed visual teaching materials on business mathematics and statistics. She was invited and talked in many tutorials such as one titled "Visually Do Statistical Shape Analysis!" in DSAA 2017 (Data Science and Advanced Analytics).