Scientific Research Publishing

Journal of Mathematical Finance, 2017, 7, 896-907
http://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

Lexicon Creation for Financial Sentiment Analysis Using Network Embedding

Ryo Ito[1], Kiyoshi Izumi[1], Hiroki Sakaji[1], Shintaro Suda[2]

[1]School of Engineering, the University of Tokyo, Tokyo, Japan
[2]Mitsubishi UFJ Trust Investment Technology Institute Co. Ltd., Tokyo, Japan
Email: m2016rito@socsim.org, izumi@sys.t.u-tokyo.ac.jp, sakaji@sys.t.u-tokyo.ac.jp, suda@mtec-institute.co.jp

How to cite this paper: Ito, R., Izumi, K., Sakaji, H. and Suda, S. (2017) Lexicon Creation for Financial Sentiment Analysis Using Network Embedding. *Journal of Mathematical Finance*, **7**, 896-907.
https://doi.org/10.4236/jmf.2017.74048

Received: September 15, 2017
Accepted: November 14, 2017
Published: November 17, 2017

Abstract

In this study, we aim to construct a polarity dictionary specialized for the analysis of financial policies. Based on an idea that polarity words are likely located in the secondary proximity in the dependency network, we proposed an automatic dictionary construction method using secondary LINE (Large-scale Information Network Embedding) that is a network representation learning method to quantify relationship. The results suggested the possibility of constructing a dictionary using distributed representation by LINE. We also confirmed that a distributed representation with a property different from the distributed representation by the CBOW (Continuous Bag of Word) model was acquired and analyzed the differences between the distributed representation using LINE and the distributed representation using the CBOW model.

Keywords

Polarity Dictionary, Distributed Representation, Financial Domain, Network Representation, LINE, CBOW

1. Introduction

In finance and economic field, non-structural data such as textual data have attracted attentions, and many researches recently have applied text mining to analysis of financial markets or economic phenomena. In those studies, text mining is used in order to extract from textual data the information about the market and company that do not explicitly appear in numeric data. There are several studies on network analysis of financial markets using numerical data [1] [2]. However, text data contains information that is difficult to quantify such as

international politics and news, and can reflect recent events more quickly. A sentiment score is one of extracted information from textual data, and it represents the polarity (positive or negative attitude) to a certain phenomenon. There are many researches, such as Bollen and Huina [3], which analyzed the relationship between sentiment scores and market trend.

Financial text-mining is a computerized method of automatically extracting valuable and useful information on investments from vast amounts of textual data, such as those in news articles, social networking sites (SNSs), and Tweets. The use of an appropriate dictionary, especially a polarity dictionary containing positive and negative expressions within financial contexts, is important in this area. A polarity dictionary is a dictionary that assigns a polarity value to each word in a form that awards a plus polarity value to a word having a positive meaning and a minus polarity value to a word having a negative meaning. A polarity dictionary is usually created manually. However, no manually created polarity dictionaries include sufficient information on economic terms or net slang. Some studies have tried to automatically create polarity dictionaries within financial contexts [4] [5].

Sentiment analysis studies are classified into a machine learning approach including Pang *et al*. [6], and a lexicon based approach including Turney [7] Kumar Ravi *et al*. [8]. In the machine learning approach, the relationship between features of textual data and a polarity is learned by the machine learning method. Then, the polarity of a new-coming text is estimated by applying a learned model. In the lexicon based approach, the polarity of the whole text is estimated by the difference between the appearance ratio of positive words and that of negative words.

In the lexicon base approach, a polarity dictionary that consists of pairs of polarity words and their polarity value is required. It is however difficult to decide a polarity value manually to a huge number of words. Moreover, since the polarity of a word depends on the background and context of a text, the polarity dictionary should reflect them. For example, negative words in the general polarity dictionary H4N (Harvard-IV-4 TagNeg) sometimes do not have a negative polarity in the context of a finance [9]. Thus, the sentiment analysis of financial text requires a polarity dictionary specialized for a finance domain. For these reasons, the automatic construction of a polarity dictionary is critical for the lexicon base approach.

Under such a background, some studies, such as Jegadeesh and Wu [10], used text mining to evaluation of the effect of financial policies. Jegadeesh and Wu [10] applied LDA to the extraction of topics from the minutes of Federal Open Market Committee (FOMC), which is a committee that decides a U.S. financial policy. Next, they evaluated the sentiment of each topic using the polarity dictionary of the financial domain, Loughran and McDonald [9]. Finally, they analyzed the influence of the sentiment of each topic to macro variables or asset prices. However, the dictionary of Loughran and McDonald [9] was made based

on the financial report of a company, and it is not specialized for the field of a financial market. Ito *et al*. [11] analyzed the relation between a market participant's expectation formation and the sentiment of each topic obtained based on the method of Jegadeesh and Wu [10]. Although their dictionary was specializing in analysis of a financial market, it was created manually. Therefore, the polarity words in the dictionary were only a part of words that appeared in the text.

The purpose of this research is the automatic construction of the polarity dictionary that is specialized for the analysis of a financial policy. In order to express features of words quantitatively, we propose the automatic construction method of the dictionary using the network representation progressing in recent years.

2. Construction of Polarity Dictionary Using Network Representation Learning

In this section, we propose a method of constructing a Polarity dictionary using network representation learning. Firstly, we will describe the framework of our method and explain the details of each step of our method.

2.1. Framework of Proposed Method

Our method takes into consideration the relationship between polarity words and the other words that appear in each sentence. Figure 1 shows examples of dependency structures of a polarity word and a word that the polarity word depends on.

Note that all the polarity words "increased", "decreased", "dropped", and "surged" depend on the same word "rate" in these examples. Although "increased" and "decreased" are antonyms, but they are classified into the same category, words that are related to sizes and volumes. Thus, they may depend on common words. And their synonyms such as "dropped" and "surged" tend to

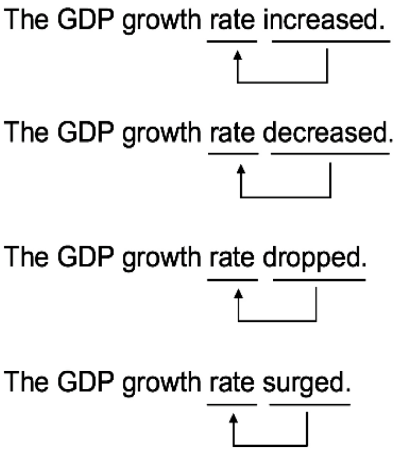

Figure 1. Examples of a polarity word and a word that the polarity word depends on.

depend on common words. From the above discussions, the dependency relation between words may reflect features of polarity words.

In this study, we used the network representation in order to present dependency relations between words. Consider the network representation of the dependency relations in **Figure 1**. On the network, the polarity words are connected to the same word "rate", and they have a relation of secondary proximity. If the secondary proximity in a dependency network is indexed, we can define the probability of being a polarity word. Firstly, we constructed dependency networks from documents about financial policy. Secondly, we qualified each node (word) of the network using LINE (Large-scale Information Network Embedding) method considering the secondary proximity. LINE is a learning method of network representations proposed by J Tang *et al*. [12]. LINE improves learning accuracy of node representation more than previous methods such as Deep Walk [13] by using information of secondary proximity. Moreover, LINE can easily handle large-scale networks with millions of vertices and edges. Finally, we tried to construct a polarity dictionary using the bootstrapping method, the distributed representation calculated by LINE method, and small number of seed polarity words.

Katakura and Takahashi [14] also used distributed representations in order to construct a polarity dictionary. They used CBOW (Continuous Bag of Word) model [15] for the calculation of the distributed representations. The CBOW model uses the information about neighbors around each word to predict the current word from a window of surrounding context words. The order of context words does not influence prediction. Since the variation of neighbor words is huge, the distributed representation using CBOW tends to have broader information than that using LINE. Since LINE focuses on the dependency relations, it is expected that the distributed representation using LINE is better fitted to determination of polarity words.

2.2. Construction of Dependency Network

As the first step, we constructed a dependency network between words from the text to be analyzed. Since we try to construct a dictionary specialized for the evaluation of monetary policy, we used FOMC minutes that is the minutes of the committee to formulate the monetary policy of the United States.

Firstly, we will collect HTML files of the FOMC minutes from the website of FRB (https://www.federalreserve.gov/) using web crawler program. Next, we parsed the HTML files and extract only the text part. FOMC minutes consists of several sections and we analyzed the following 5 sections that are denoted the review of economic situation and outlook: Developments in Financial Markets and Open Market Operations; Staff Review of the Economic Situation; Staff Review of the Financial Situation; Staff Economic Outlook; Participants' Views on Current Conditions and the Economic Outlook; Committee Policy Action. Then, by performing dependency analysis on the acquired text, dependency

pairs of words are extracted. At this time, lemmatization is performed to each word. Furthermore, we create a weighted directed graph from word dependency pairs. Each node of the network corresponds to each word, and a weight w_{ij} of an edge linked from a node v_i to another node v_j corresponds to a frequency of the dependency pair of the words. The direction of each edge stands for the direction of the dependency relation between words

2.3. Distributed Representation of Words Using LINE

Using the dependency network between words obtained in the previous section, we calculated the distributed representation of words by LINE, a network representation learning method. There are two methods of LINE: primary LINE and secondary LINE. We used secondary LINE because we want to acquire a distributed representation considering a secondary proximity.

In the secondary LINE, if there are many nodes which are commonly adjacent between two nodes, it will learn that the two nodes have similar distributed representations. About the probability of existence of an edge from one node v_i to another node, a probability function estimated from distributed representation and an observed probability function will be closer by the learning.

We introduce a vector $\boldsymbol{u}_i$ of node v_i and a context vector $\boldsymbol{u}_i'$ of the node. The probability that edge is created from one node v_i to another node v_j estimated from distributed representation by the following equation.

$$p_2\left(v_j \middle| v_i\right) = \frac{\exp\left(\boldsymbol{u}_j'^{\mathrm{T}} \cdot \boldsymbol{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\boldsymbol{u}_k'^{\mathrm{T}} \cdot \boldsymbol{u}_i\right)}$$

The probability that an edge is linked from one node v_i to another node v_j that can be observed from the network G having V as a set of nodes and E as a set of edges is as follows.

$$\widehat{p_2}\left(v_j \middle| v_i\right) = \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}}$$

$N(i)$ represents a set of nodes with edges linked from node *i*.

Here, the learning is performed so as to minimize the weighted sum of the distances between the above two probability distributions. The objective function is expressed by the following equation.

$$O_2 = \sum_{i \in V} \lambda_i d\left(\widehat{p_2}\left(\cdot \middle| v_i\right), p_2\left(\cdot \middle| v_i\right)\right)$$

λ_i represents the degree of importance of node *i* in the network, and its value is $\sum_{k \in N(i)} w_{ik}$. The distance $d(\cdot|\cdot)$ between two distributions is used as Kullback-Leibler divergence. By removing the constant term, the following equation is derived.

$$O_2 = -\sum_{(i,j) \in E} w_{ij} \log p_2\left(v_j \middle| v_i\right)$$

By optimizing this objective function using negative sampling, we can obtain a

distributed representation for each node.

2.4. Acquisition of Polarity Words by Bootstrapping Method

Next, we extracted candidates of polarity words by bootstrapping method with the distributed representation of words obtained by LINE.

Firstly, we manually prepared a small size of polarity words as a seed of the bootstrapping method. The seed words were added to the polarity word candidate list C. Reliability P_d is a value that indicates the word d seems to be a polarity word. The reliability of all seed polarity words is one.

Next, for each word that is not included in the polarity word candidate list, the reliability P_d is calculated by an average of a product of a similarity between word d and each word c in the polarity word candidate list and the reliability P_c of the word c.

$$P_d = \frac{1}{|C|}\sum_{c\in C} sim(\boldsymbol{u}_c, \boldsymbol{u}_d) P_c$$

$sim(\boldsymbol{u}_c, \boldsymbol{u}_d)$ is a function that stands for the similarity of distributed representations between two words, and it uses cosine similarity. $\boldsymbol{u}_d$ corresponds to a vector representation of a node (word d) obtained by LINE.

$$sim(\boldsymbol{u}_c, \boldsymbol{u}_d) = \frac{\boldsymbol{u}_c \cdot \boldsymbol{u}_d}{|\boldsymbol{u}_c||\boldsymbol{u}_d|}$$

After calculating the reliability P_d for each word, L words with the highest values of reliability P_d are added to the polarity word candidates list C.

By repeating the above steps, the polarity word candidate list C is expanded. After M times of repeating the above procedure, our system output the top N words with the highest values of reliability P_d in the polarity word candidates list C as the final candidates of polarity words.

3. Experiment Methods

In this section, we describe various experiment settings and evaluation methods.

3.1. Visualization of Distributed Representation

We visualized the distributed representation to examine characteristics of the distributed representation obtained by LINE. Especially, we focused on words near polarity words.

We used 190 documents of FOMC minutes published from January 1993 to November 2016 to construct dependency networks of words. The number of dimensions of distributed representation was set to 50. For the visualization of the distributed representation, we reduced the dimension using the t-SNE method [16] and visualized each word by placing it on the coordinate space.

As a comparative experiment, we also visualized distributed representation obtained by CBOW using the t-SNE method. When using the CBOW model, the number of dimensions of the distributed representation is assumed to be 50 di-

mensions like LINE and the window size in the CBOW model is 4.

3.2. Polarity Dictionary Construction Test

We conducted another experiment to acquire polarity words by carrying out bootstrapping method using distributed representations obtained by LINE. As a comparative experiment, we verified what kinds of words are acquired compared with the case of using distributed representations obtained by CBOW model as input. The parameters of in the bootstrap L, M, and N were set to 1, 30, and 30, respectively. The seed polarity words are "increase", "high", and "improvement". The distributed representations in LINE and CBOW model both used the distributed representation obtained in the previous section.

4. Results and Discussion

In this section, we will describe experimental results. In particular, we discuss differences between distributed representation by LINE and that by CBOW from the viewpoint of obtaining polarity words.

4.1. Visualization of Distributed Representation

Figure 2 is a mapping of the distributed representation of words obtained by LINE on a two-dimensional plane. As can be seen, the mapped distributed representation is divided into upper and lower parts. The words in the upper part include words, such as adverbs and numbers that are likely to be depended from the other words. In other words, the distributed representation by LINE reflected the dependency relationship of words.

Figure 3 is an enlarged view of a part of Figure 2. Many polarity words placed closely such as "increase", "decrease", "rise", and "drop". Also in the other parts,

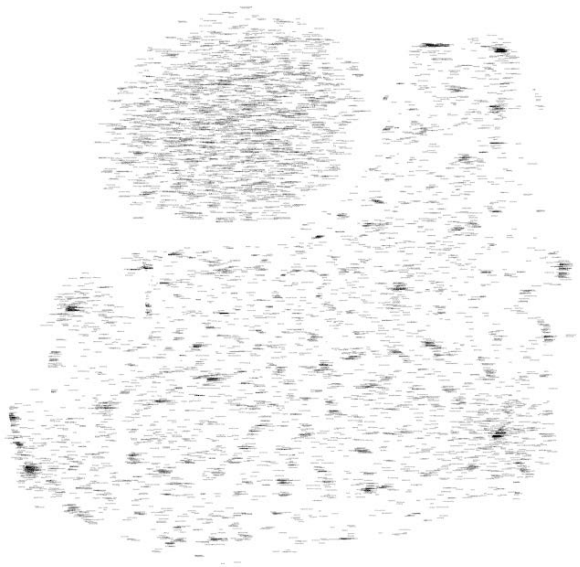

Figure 2. Visualization of words' distribution representation by LINE.

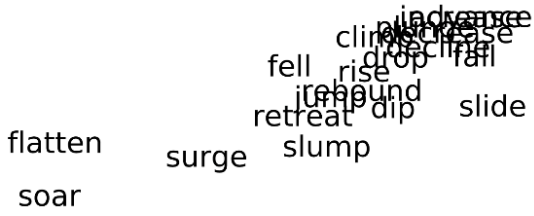

Figure 3. Visualization of polarity words' distribution representation by LINE.

there were clusters of noun polarity words or adjective polarity words. These results show the possibility of automatically acquiring polarity words by giving seed words based on the relationship of secondary neighbors in the dependency network.

Figure 4 is a mapping of the distributed representation of words obtained by the CBOW model on a two-dimensional plane. As can be seen from the figure, there was no clear separation between the two parts compared to LINE. Figure 5 is an enlarged view of a part of Figure 4. As can be seen, many polarity words existed concentratedly, including "increase", "decrease", "growth", and "expansion". This indicates the possibility of automatically acquiring polarity words by giving seed words even in case of using CBOW model.

In LINE's case, polarity words were separated in accordingly to their part of speech. On the other hand, in CBOW's case, it was found that the polarity words tend to be distributed collectively regardless of parts of speech to some extent compared to LINE. That is because LINE uses grammatical structure in order to acquire distributed representation based on dependency relations. In contrast, CBOW uses neighborhood information between words to calculate distributed representation, and distributed representation can include semantic information.

4.2. Polarity Dictionary Construction Test

Table 1 shows 20 words with the highest reliability among the words acquired by using the bootstrapping method with the distributed representation by each model as input.

In the case of using distributed representation by LINE, it can be seen that polarity words such as "decrease", "drop", "advance", "fall", and so on are acquired. In particular, words such as "climb" do not have polarity in the usual context, but in the sense that prices rise, they are popular representations in financial context. The proposed method was able to acquire such polarity words specialized in the financial domain. It however acquired some wrong words like "there" as polarity words.

Also in CBOW's case, we could acquire polarity words such as "decline", "rise", and "drop". It could also acquire specialized polarity words such as "climb" for the context of monetary policy.

We will describe the difference between LINE's case and CBOW's case. In

Figure 4. Visualization of words' distribution representation by CBOW.

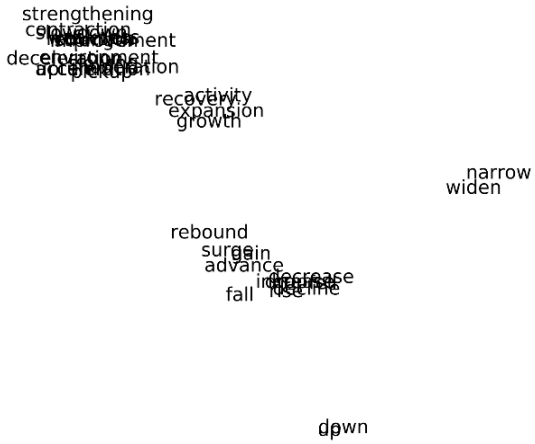

Figure 5. Visualization of polarity words' distribution representation by CBOW.

LINE's case, semantic drifts [17] occurred at earlier step of the bootstrapping method. The semantic drift is a problem that a bootstrapping method extracts wrong instances during iterative processing. For an instance, after extracting a non-polarity word "differently", the bootstrapping method extracted another wrong word "differing". The suppression of semantic drift is one of future works.

5. Summary

In this study, we aim to construct a polarity dictionary specialized for the analysis of financial policies. Based on an idea that polarity words are likely located in the secondary proximity in the dependency network, we proposed an automatic dictionary construction method using secondary LINE that is a network representation learning method to quantify relationship.

Table 1. Polarity words obtained by bootstrapping method with LINE and CBOW.

Step	LINE	CBOW
1	rise, decline, decrease,	decline, rise,
2	drop, advance, climb,	drop, decrease,
3	showing, fall,	advance, fall,
4	jump, there,	fell, surge,
5	faster-than-anticipated,	jump, decelerate,
6	quickened, constraining,	climb, rebound,
7	weaker-than-anticipated,	contract,
8	slower-than-expected,	accelerate,
9	differently,	deceleration,
10	better-than-anticipated,	run-up, gain,
11	differing,	step-up,
12	higher-than-expected,	weaken,
13	approximating	recover

The results suggested the possibility of constructing a dictionary using distributed representation by LINE. We also confirmed that a distributed representation with a property different from the distributed representation by the CBOW model was acquired and analyzed the differences between the distributed representation using LINE and the distributed representation using the CBOW model.

There are some tasks to be addressed in the future, but the following four points are mainly listed as important issues.

The first point is to quantitatively evaluate the accuracy of dictionary constructed by the proposed method. In addition, it is necessary to compare the accuracy of using the CBOW model and the accuracy of using the proposed method. It is also necessary to verify the robustness of a polarity dictionary depending on kinds of seed words.

The second point is the determination of sentiment in polarity dictionary. The proposed method added a new word with high reliability to the polarity word candidate list regardless of the positive/negative of the word. In order to construct a dictionary classified as positively/negative of acquired words, however, the proposed method should be extended. To do so, it is necessary to obtain a distributed representation considering synonyms and antonym relations of words. K. A. Nguyen *et al*. [18] considered a synonym/antonym relation in the objective function of the Skip-gram model. Also in this research, by extending LINE based on this method, it is possible to obtain distributed representations considering the synonyms and antonym relations of words, and to construct a dictionary in which the acquired polarity words are divided into positive and negative.

As the third point, when the bootstrapping method is used, it is necessary to avoid semantic drift in which words other than polarity words are acquired

during the iteration proceeds. To do so, we should redesign the function of reliability for each word, and/or integrate a bootstrapping method with a learning method. It would be effective to restrict word extraction using a list of words that may cause semantic drift [17].

Finally, we should test the effectiveness of the obtained polarity dictionary for sentiment analysis like Ito *et al*. [11]. We can check the accuracy of the obtained dictionary using the accuracy of sentiment analysis. Performance of sentiment analysis with the obtained dictionary is compared with that without the obtained dictionary. Furthermore, if the accuracy of sentiment analysis is improving, we will examine the relationship between sentiment scores obtained and some economic indices.

Statement

The views expressed in this paper are those of the authors and do not necessarily reflect the official views.

References

[1] Bollen, J. and Huina, M. (2011) Twitter Mood as a Stock Market Predictor. Computer, **44**, 91-94.

[2] Curran, J.R., Murphy, T. and Scholz, B. (2007) Minimising Semantic Drift with Mutual Exclusion Bootstrapping. *In Proceedings of the* 10*th Conference of the Pacific Association for Computational Linguistics*, University of Melbourne, Australia, Sep 19-21 2007, 172-180.

[3] Ito, R., Suda, S. and Izumi, K. (2017) Impact Analysis of Forward Guidance on Market Expectation—Text Mining Approach—46*th Winter JAFEE Competition of* 2016, Musashi University, Tokyo, Japan, Feb 17-18 2017, 60-71.

[4] Izumi, K., Suzuki, H. and Toriumi F. (2017) Transfer Entropy Analysis of Information Flow in a Stock Market. In: Aruka, Y. and Kirman, A., Eds., *Economic Foundations for Social Complexity Science: Theory, Sentiments and Empirical Laws*, Springer, Berlin.

[5] Jegadeesh, N. and Wu, D. (2015) Deciphering Fedspeak: The Information Content of FOMC Meetings, 2016 AFA Anual Meeting Working Paper. San Francisco Marriott Marquis, San Francisco, USA, Jan 03-05 2016. https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=1136

[6] Kamps, J., Marx, M., Mokken, R.J. and de Rijke, M. (2004) Using WordNet to Measure Semantic Orientations of Adjectives. *In Proceedings of the* 4*th International Conference on Language Resources and Evaluation*, Universidade Nova de Lisboa, Lisbon, Portugal, May 26-28 2004, 1115-1118.

[7] Katakura, K. and Takahashi, O. (2015) Dictionary Preparation and Financial Market Analysis by Distributed Representation Learning of Financial Market News. *The* 2015 *National Conference of Artificial Intelligence* (No. 29), Hyatt Regency Austin, Texas, USA, Jan 25-30 2015.

[8] Loughran, T. and McDonald, B. (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, **66**, 35-65.

[9] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv: 1301.3781.

[10] Maaten, L.V.D., and Hinton, G. (2008) Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.

[11] Nguyen, K.A., Walde, S.S.I. and Vu, N.T. (2016) Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Berlin, 7-12 August 2016, 454-459. https://doi.org/10.18653/v1/P16-2074

[12] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, July 2002, 79-86.

[13] Perozzi, B., Al-Rfou, R. and Skiena, S. (2014) DeepWalk: Online Learning of Social Representations. *Proceedings of the* 20*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 701-710. https://doi.org/10.1145/2623330.2623732

[14] Turney, P. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics ACL*'02, Philadelphia, July 2002, 417-424.

[15] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q. (2015) LINE: Large-Scale Information Network Embedding. *Proceedings of the* 24*th International Conference on World Wide Web*, Florence, 18-22 May 2015, 1067-1077. https://doi.org/10.1145/2736277.2741093

[16] Tsubouchi, K. and Yamashita, T. (2014) Generation of Positive Negative Dictionary for Finance Using Stock Bulletin Board Data. National Convention of Artificial Intelligence.

[17] Xu, R., Wong, W.-K., Chen, G. and Huang, S. (2017) Topological Characteristics of the Hong Kong Stock Market: A Test-Based P-Threshold Approach to Understanding Network Complexity. *Scientific Reports*, **7**, Article ID: 41379. https://doi.org/10.1038/srep41379

[18] Yanagimoto, H. (2014) Improvement of Sentiment Dictionary Using Neural Network Language Model. *Proceedings of the* 28*th Annual Conference of the Japanese Society for Artificial Intelligence*, Himegin Hall, Ehime, Japan, May 12-15 2014.