IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 14, 2020, accepted August 27, 2020, date of publication September 2, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021182

Detecting a Risk Signal in Stock Investment Through Opinion Mining and Graph-Based Semi-Supervised Learning

BYUNGUN YOON, (Senior Member, IEEE), YUJIN JEONG, AND SUNHYE KIM
Department of Industrial and Systems Engineering, Dongguk University, Seoul 100715, South Korea

Corresponding author: Byungun Yoon (postman3@dongguk.edu)

This work was supported by the National Research Foundation of Korea under Grant 2019R1A2C1085388.

ABSTRACT The objective of this study is to develop an algorithm to support a decision-making process in stock investment through opinion mining and graph-based semi-supervised learning. For this purpose, this research addresses the following core processes: (1) filtering fake information, (2) assessing credit risk and detecting risk signals, and (3) predicting future occurrences of credit events through sentiment analysis, word2vec, and graph-based semi-supervised learning. First, financial data, including news, texts in social network services, and financial statements, were collected. Among these data, fake information such as rumors and fake news was filtered by author analysis and a rule-based approach. Second, credit risk was assessed by opinion mining and sentiment analysis for both social data and news in the form of a sentiment score and the trend of documents for each stock. A signal for a credit event was then detected by the degree of assessed risk. Consequently, the possibility of credit events such as delisting and bankruptcy in the near future was forecast based on the risk signal using logistic regression. This research illustrated the real case of a company to validate the applicability of the proposed approach. The results of this study can help investors monitor a large amount of historically accumulated data and detect hidden signals of risk events ahead of time.

INDEX TERMS Decision support system, early signal detection, graph-based semi-supervised learning, logistic regression, opinion mining.

I. INTRODUCTION

A global economic crisis arose from the subprime mortgage crisis of 2008, resulting in corporate bankruptcies or delistings from securities markets amid a shrinking national economy [1]. Such an economic crisis is usually caused by an accumulation of small events leading to a potentially great impact [2]. If these small events can be recognized and caught beforehand, a severely damaging national or global crisis may be prevented by treating them ahead of time. Recently, there have been many uncertain phenomena, such as cryptocurrency (bitcoin), a newly emerging financial service. Protectionism, which is dispersed worldwide, causes trade wars between countries, and furthermore, it may lead to a slowdown of economic growth in the long term. Thus, monitoring and prevention based on early detection have been emphasized in recent times as important for avoiding credit events that might cause a national and global economic crisis and for minimizing socioeconomic losses.

Despite the increased significance of early detection through the monitoring of previous data, it was difficult to collect various types of data before 'the era of information revolution' came. Previous prediction models were mostly developed in the form of a one-dimensional model, leading to inaccurate results. The recent era of information with its frequent use of information technology (IT) devices has increased the amount of data as well as the availability and accessibility of a variety of databases, such as social network services (SNSs) and web communities. This increase of data enables not only experts and analysts but also individual investors to obtain superior-quality financial and non-financial data about firms. In particular, social data can provide information about investor behavioral patterns and the organizational behaviors known to affect fluctuations in stock prices [3]–[5]. Thus, historically accumulated data

The associate editor coordinating the review of this manuscript and approving it for publication was Chaitanya U. Kshirsagar.

can be a critical source to catch movements in financial markets [6].

Since the importance of monitoring and detecting signals has been highlighted, there have been attempts to detect risk signals or bankruptcy by considering various factors that could affect the risk of company failure [7]–[11]. These studies have deduced that diverse factors (internal and external factors, financial factors, noneconomic factors, corporate culture, and the attitudes of managements and investors) can influence firms' potential for failure by intimately interacting with each other. In addition, with the development of natural language processing (NLP), there have been attempts to use text data for financial analysis, and it has developed into a new field called natural language-based financial forecasting (NLFF) [12]–[15]. However, these studies have some limitations. First, most studies have focused on predicting increases or decreases in stock prices, but attempts to detect an early signal of credit events have rarely been made [16]–[19]. Second, there have been many studies using numerical data, mainly stock prices and financial statements, and there is a dearth of research using social data. Recently, there have been studies using news titles to detect events affecting business, but the results have not led to stock investment decision-making [20], [21]. In addition, in most studies, some factors influencing corporate bankruptcy, such as customer behavior, were still not considered. Last, in the case of companies not listed on securities markets, there is no available information since such companies are not obliged to provide their information to the public. Therefore, there is a lack of research on these private companies.

Therefore, the objective of this study was to develop an algorithm to support decision-making in stock investment using both objective and subjective databases through opinion mining and machine learning. The algorithm developed in this study involved the following three parts: (1) data collection and filtering, (2) credit risk assessment and early signal detection, and (3) prediction of credit events. First, data were collected from various databases related to stock investment ranging from news and financial statements to social network services and web communities. Fake information, such as rumors and fake news, was then filtered by author analysis and a rule-based approach. Second, a risk signal, which is a sign or trigger of credit events such as bankruptcy and delisting, was detected from sentiment analysis and opinion mining. The risk signal was defined by three grades ('dangerous', 'warning', and 'caution') in stock investment to provide insights for monitoring and responding to credit events in advance. Third, the possibility of credit events occurring was predicted by logistic regression, including the dependent variable in the form of binary values (occurring or not occurring) and independent variables based on the results of signal detection.

The suggested algorithm focuses on detecting hidden signals and predicting the possibility of credit events in advance to support decision-making when investors trade or manage their stocks. Although the financial status presented in financial statements may seem stable, some events that can result in the occurrence of serious credit events might be hidden or not easily recognized by investors or firms. For that reason, this paper attempts to find signs or triggers of future credit events earlier and help decision-makers respond to them. In this study, various sources, including financial status and opinions expressed by investors, are utilized after filtering fake information. A comprehensive analysis based on objective and subjective databases supported the credibility and feasibility of the proposed approach. Our study contributes to improving the availability of subjective data, such as the opinions of investors, to find hidden information and signals in advance without information related to financial status, such as the monitoring of private equity. This algorithm contains a unique process for dealing with fake information, which may have an impact on investing in stocks. The proposed algorithm using graph-based semi-supervised learning and a logistic regression model contributes to more accurately detecting signals and predicting future risk events in the stock market. Graph-based semi-supervised learning propagates information from labeled to unlabeled data using similarity, which helps estimate the sentiment value of words while considering context in stock markets.

This paper is organized as follows. Section 2 describes the background literature related to opinion mining and graph-based semi-supervised learning. In section 3, the overall research framework and process are explained. In section 4, two case studies using the proposed algorithm are described. Based on the case studies, section 5 discusses practical implications. Section 6 summarizes the overall research with academic implications and possibilities for future research.

II. LITERATURE REVIEW

A. OPINION MINING

Opinion mining is a method for analyzing opinions by collecting and processing documents from various datasets [22]–[24]. Opinion mining analyzes the opinions, evaluations, attitudes, and emotions of people who express their impressions and thoughts about specific topics posted on social networks and websites [25], [26]. The sentiment analysis used in this paper analyzes emotions expressed by people as a subfield of opinion mining and has recently been used in many fields of application, including the financial field [27]–[29]. Opinion mining can deal with a large number of documents in the form of text or scores [30], [31]. Opinion mining has been applied in a wide range of fields [32]–[34]. The general process for conducting opinion mining research consists of three phases. First, a sentiment dictionary in which each word has sentiment value is constructed by the conjunction method, the pointwise mutual information (PMI) method, the WordNet exploring method, the gloss classification method, evaluation theory, natural language processing (NLP), and statistical schema matching [35]. In the sentiment dictionary, a numerical value represents the degree of sentiment for each word. Second, the sentiment value is propagated from labeled to unlabeled words through the PMI, machine learning, and NLP combined methods. Then, all opinions are summarized and explained [36]. Establishing sentiment dictionaries based on the well-defined opinions of authors is the most important step in performing accurate analysis. This study focuses on finance, which is appropriate for applying opinion mining because the field has an obvious relationship between sellers and buyers who write relatively clear words [37].

As mentioned earlier, increases in the use of SNSs and platforms for sharing information continue to extend the amount and range of available data. On the other hand, this trend encourages us to distinguish between correct and fake information because fake information or excessive data may cloud the judgment of decision-makers. For this reason, filtering fake information within a large database has received much attention in recent years [38], [39]. Filtering fake information is considered a difficult task, not a simple problem, because the filtering of fake data involves a complex mixture of linguistics, psychology, and machine learning [40]. There have been many attempts at filtering hoaxes. Most attempts depend on the linguistic rule-, author credibility-, user-, and topic-based models. The linguistic rule-based model is the simplest of those that have been widely utilized. For the purpose of evaluating the reliability of new information, this model exploits the common linguistic features of documents with low credibility, such as vocabulary and syntax [41], [42]. The author-based model examines whether the author is a spammer who writes false information. If authors are spammers, their information is classified as fake information [43]–[48]. The user-based model is widely used today because of SNSs. This model uses an index scored by the user, who accesses the information as a learning label [49]–[51]. Finally, the topic-based model is the way that big data enables us to distinguish an opinion as true if it is involved in most opinions related to a specific topic, as opposed to false opinions, which constitute a minority of overall opinions [52], [53].

B. GRAPH-BASED SEMI-SUPERVISED LEARNING

Semi-supervised learning has considerably better efficiency and accuracy than the abovementioned techniques since the number of independent variables is sufficient while the number of dependent variables is insufficient (a general situation) [55]. Therefore, when sufficient input words can be obtained, compliance learning can be useful in the calculation of emotional scores that are difficult to obtain from the emotional value of the words that have been printed. In social data analysis, although much text is generated due to active user interactions, the sentiments of most words are not presented and need to be calculated by using the sentiment values of several words that are already known. For this reason, there are many studies using semi-supervised learning to analyze social data [56]–[58]. Figure 1 shows a typical example of image recognition when semi-supervised learning is widely applied. Graph-based semi-supervised learning depends on the assumption that if two nodes are connected by strong edges on a graph consisting of nodes and edges, then the labels of these two nodes tend to be similar. Graph-based semi-supervised learning algorithms include the harmonic function algorithm [59] and manifold regularization algorithm [60]. In our study, we used sentiment value propagation based on the basic assumption of graph-based semi-supervised learning. In the proposed approach, while nodes are words extracted from texts, edges refer to weighted labels based on similarities between nodes. First, words are defined as independent variables, while sentiment values obtained from opinion mining are defined as dependent variables. Using a vector of words extracted from word2vec, the cosine similarity used as a weight is then calculated. Finally, weighted labels are propagated to the unlabeled data.

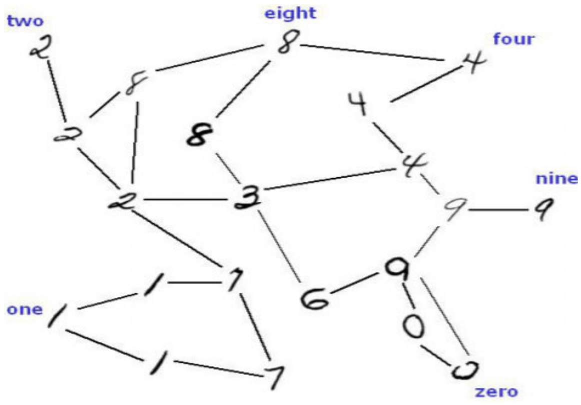


FIGURE 1. An example of graph-based semi-supervised learning.

III. METHODOLOGY

A. BASIC CONCEPT

To support the decision-making process in stock investment, we propose a new algorithm to detect risk signals and predict the future occurrence of credit events. A risk signal is defined as a warning that investors should pay attention to the status and management issues of firms when they are concerned about dealing in those companies' stocks. The risk signal can be detected by estimating the sentiment value of data, including news and opinions, using sentiment analysis based on opinion data, word2vec, and graph-based semi-supervised learning. This is because the stock price is decided by the business activities of firms and investment behaviors of investors presented as opinions or reviews posted on websites or social network services. In addition, critical information related to firms might be hidden in the opinions of analysts and other investors. This information can serve as influential evidence when making decisions to buy or sell stocks. The possibility of credit events is then predicted by a logistic regression model that is composed of indicators based on the sentiment value of opinions. In this paper, a credit event is

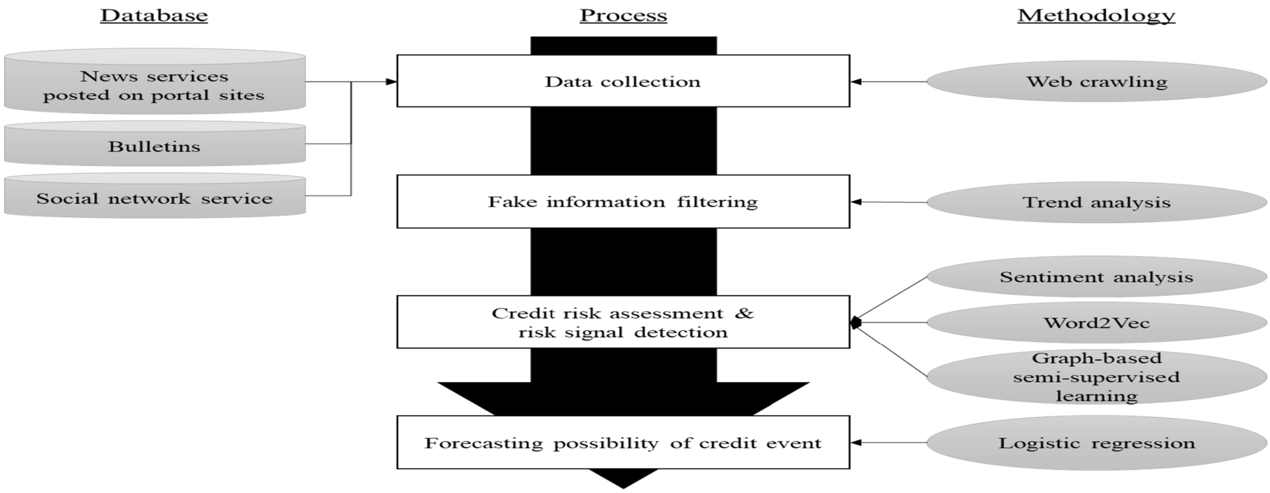

Database
Process
Methodology
News services posted on portal sites
Bulletins
Social network service
Data collection
Fake information filtering
Credit risk assessment & risk signal detection
Forecasting possibility of credit event
Web crawling
Trend analysis
Sentiment analysis
Word2Vec
Graph-based semi-supervised learning
Logistic regression

FIGURE 2. Research framework.

defined as an incident that seriously affects the bankruptcy risk of a company, such as default, court receivership, and rehabilitation procedures. These events can be identified in financial statements and public announcements that are disclosed to the public with regard to critical issues for enterprise management. To develop a logistic regression model for forecasting future occurrences, data of firms involved in the same industry, such as the sentiment value of opinions derived from a prior phase, are utilized. The detailed process is shown in Figure 2 and is explained in the next section.

B. OVERALL PROCESS

1) DATA COLLECTION AND INFORMATION FILTERING

This paper utilizes all information related to stock investment obtained from news and opinions posted on websites and SNSs. The news is written based on facts and is regarded as quite objective, as there are few inferences based on facts. The news is provided to investors through websites as well as papers published by the press and can be collected from each portal site and press. Since portal sites mostly provide news services with limited financial news by extension, it is useful to collect articles related to investment rather than using individual, general press articles. These opinions are also collected from finance websites, online communities, and SNSs. A private investor is apt to depend on information posted on diverse communities because hidden information can help one decide whether to invest in certain stocks. After investigating website candidates for collecting news and opinion data in advance, the final database is chosen according to its availability, credibility, degree of subjectivity, and volume.

After collecting raw data from web services, these data need to be refined to clarify documents and derive more accurate results. Recently, many rumors and fabrications have been used to instigate or manipulate stock prices with the intent to generate additional margin. Thus, recent studies have attempted to resolve how to filter spam or fake opinions by three approaches: behavior-, language-, and graph-based. This paper suggests behavior- and language-based approaches in sequence, which means that different characteristics of language in fake or genuine opinions are inspected by sentiment analysis after examining the current trend of opinion occurrences. The behavior-based approach aims to identify the distribution of opinions by investors, while the language-based approach can pinpoint the pattern of opinions.

At first, abnormal symptoms—i.e., abnormal movement such as rapid growth in the number of opinions—are discovered by analyzing the trend in articles, including news and social data. For the purpose of inciting investors to buy or sell stocks, most fake information occurs during a certain period. The more prevalent an opinion is, the easier it is for investors to be influenced by fake information. People who instigate investment by others may acquire margin from its associated effects. Hence, this paper preferentially discovers the time period showing the most drastic growth. Second, the characteristics of opinions occurring during the period that shows a steep rise are then investigated for similar linguistic patterns. If specific linguistic phrases or expressions are iterated, they can be regarded as fake information or hoaxes created for purposes of instigating actions by other investors.

2) CREDIT RISK ASSESSMENT AND RISK SIGNAL DETECTION

This paper attempts to propagate the sentiment value of core keywords to relevant words after allocating sentiment value for each document with naïve-Bayes classification, word2vec, and graph-based semi-supervised learning. After filtering fake information, all textual data are preprocessed by natural language processing to assess credit risk. Data preprocessing includes the following processes: (1) splitting

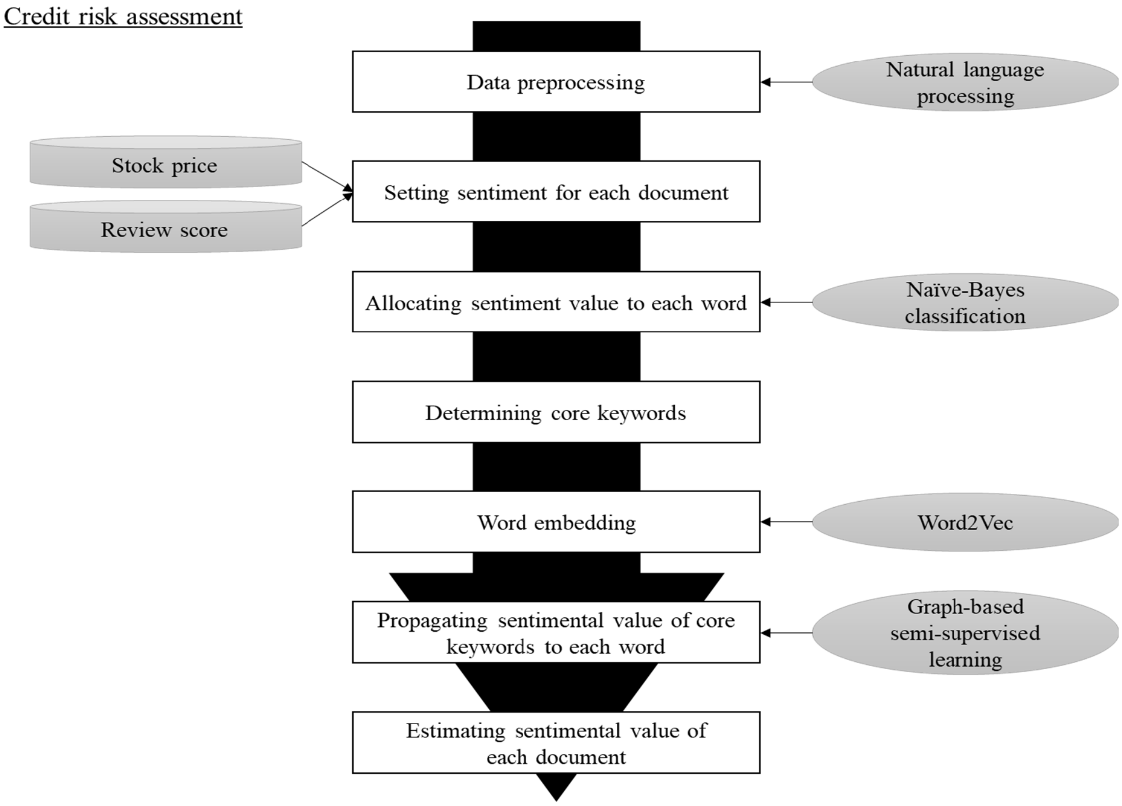
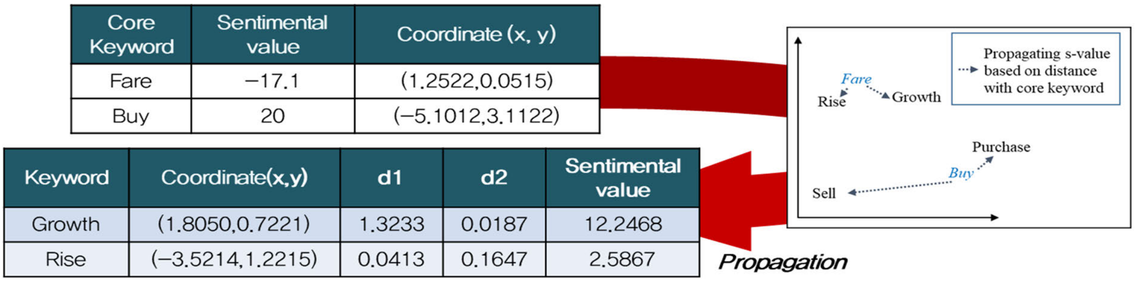

Credit risk assessment
Data preprocessing
Natural language processing
Stock price
Review score
Setting sentiment for each document
Allocating sentiment value to each word
Naïve-Bayes classification
Determining core keywords
Word embedding
Word2Vec
Propagating sentimental value of core keywords to each word
Graph-based semi-supervised learning
Estimating sentimental value of each document

FIGURE 3. Process for assessing credit risk.

sentence, (2) tokenizing, and (3) part-of-speech (POS) tagging and parsing. This preprocessing aims to rate documents using the sentiment values of sentences and words. The sentiment value of each word is estimated by propagating the sentiments of relevant documents. In other words, sentiment values such as positive, neutral and negative are allocated to words presented in the document. The sentiment value has a numeric value of +1 (positive), 0 (neutral), or −1 (negative).

To estimate the sentiment value of words, the document is preferentially rated, which is conducted in two ways: stock price and review score. The rating differs depending on whether the company is listed on a securities market, such as KOSDAQ (Korean Securities Dealers Automated Quotations), or not. Information related to listed companies is disclosed to the public, and the stocks of listed companies are traded on the applicable exchange. However, the stocks of unlisted companies are traded in person. Since the stocks of exchange-listed companies have quoted prices while unlisted companies do not, the sentiment of documents related to listed companies is defined by stock price while that of unlisted companies is determined based on the score determined from the reviews provided by users.

The review score is derived from user expressions of whether they will buy or sell stocks based on subjective judgment. This score is usually rated on a five-point scale, such as strongly recommend buying (or selling) stock, cautiously recommend buying (or selling) stock, and neutral. If a user strongly recommends buying stocks on a post on a bulletin board or an SNS, then the post or review is rated as strongly positive and +1. The opposite sentiment is rated in the same way. All reviews with scores are assigned a value of +1 (positive) or −1 (negative). After deciding the sentiment of each article, this sentiment value is disseminated to all words included in the article through Naïve-Bayes classification, which is based on the co-occurrence of words in the article. This process has the advantage of allocating the sentiment value of words stochastically. Definitive values for words are deduced by aggregating all documents. Although the sentiment value of each word is defined at this step, it cannot reflect the relation between words in documents and semantic ambiguity in different fields. To alleviate this issue, core keywords with large absolute values are chosen. The value of core keywords is then propagated again in proportion to the distance between words through graph-based semi-supervised learning. This distance is derived from a word2vec-based map developed by considering relationships between words in the corpus. Word2vec is a technique for word embedding that represents a word as a vector, allowing for context in the corpus. As a result, each word with a unique value can be mapped on a two-dimensional map. The more similar the meanings of words, the closer the distance between words on the map. Based on the distance between words, the sentiment value of core keywords is propagated to words surrounding core keywords,



Core Keyword	Sentimental value	Coordinate (x, y)
Fare	−17.1	(1.2522,0.0515)
Buy	20	(−5.1012,3.1122)

Keyword	Coordinate(x,y)	d1	d2	Sentimental value
Growth	(1.8050,0.7221)	1.3233	0.0187	12.2468
Rise	(−3.5214,1.2215)	0.0413	0.1647	2.5867

FIGURE 4. The process for word propagation.

as shown in Figure 4. Likewise, the net sentiment values of words are estimated by summing the proximity indices between each word and core keywords that are calculated by considering distances between them and the sentiment values of core keywords, as shown in Figure 4 and (1), (2), (3), as shown at the bottom of the page. For this process, graph-based semi-supervised learning is conducted by propagating a part of the labeled data (sentiment value of core keyword) to unlabeled data relying on the distance on the map (it can be seen as a graph). The sentiment value of each document is finally estimated by summing the value of words involved in the document.

The credit risk for detecting a signal in advance is evaluated by an integrated indicator (called the credit risk metric) composed of sub-indices based on the sentiment value. The sentiment values for different documents calculated at the prior step are averaged by date, week, and month to develop sub-indices, as shown in Table 1. All sub-indices are defined with the purpose of detecting a signal and a warning for careful investment by investors in this paper. The first is the average of accumulated sentiment value by date. This indicator is appropriate for understanding the accumulated feelings about specific stocks at a certain point in time and can aid in avoiding the bias caused by the sudden growth and decline in sentiment value that comes from relying on the number of reviews. The second indicator is a period of sustained similar sentiment represented as days. If the sentiment continues for a certain period, then there is a high probability of credit events because of that sentiment, which means that events occurring at that time that induce positive or negative feelings from investors are not transient. These events can be regarded as a sign of economic accidents. The third is the number of changeovers from positive to negative sentiment and vice versa during a certain period. The higher the number of sentiment changeovers, the more controversial the events are. It is possible to interpret this as the presence of arguments about the pros and cons for events during this period and high interest in the stock. The final sub-index is the growth rate of sentiment value, making it possible to identify overall trends for stocks. Since all sub-indicators have different scales, such as days and percentages, they are normalized from 0 to 1 using the extreme value of each indicator. This value is then aggregated by the weighted sum and presented as a risk signal metric in this paper. This metric for detecting a signal can be classified into three grades

TABLE 1. Indicators for detecting risk signal.

Integrated indicator	Sub-index	Unit	Description
Credit Risk metric	*Average of accumulated sentiment value*	Score	• The average of accumulated sentiment value • Considering days in which there are no opinions
	Duration of sustained negative opinions	Day	• The longer that negative opinions occur consistently, the higher the possibility that the credit event happens
	The number of intersections of positive and negative opinions	Number	• The number of changes in sentiment during two weeks (from positive to negative or from negative to positive) • The higher the sentiment intersects, the more that interest in the company increases
	The change rate of daily sentiment value on average	Percentage	• The variation of sentiment by dates • Identifying trend of sentiment

$$Sentimental\ value\ of\ keyword\ i = \sum\nolimits_{j=1}^{m} (Proximity\ index\ between\ keyword\ i\ and\ core\ keyword\ j) \tag{1}$$

$$Proximity\ index\ between\ keyword\ i\ and\ core\ keyword\ j = \frac{1}{r^2} * (Sentimental\ value\ of\ core\ keyword\ i) \tag{2}$$

$$r_j = (Euclidean\ distance\ between\ kyword\ i\ and\ core\ keyword\ j) = \sqrt{/(x_i - x_j)^2 + (y_i - y_j)^2} \quad (j = 1, 2, \ldots, m) \tag{3}$$

TABLE 2. Variables for logistic regression.

	Variable		*Description*	*Relevancy of risk signal*
Dependent variable (Y)	Occurrence of credit event (y)		• Whether credit event has occurred by date • 0 or 1 (binary)	-
Independent variable (X)	The period for which the same sentiment is maintained	Positive sentiment (x_1)	• How long positive sentiment is maintained (days) • The degree to which investors or authors respond positively for stock investment	• If positive opinions occur continuously, the optimistic prospects for the stock are maintained for a long time • Investors feel that the stock is stable • The occurrence probability of a credit event is relatively low
		Negative sentiment (x_2)	• How long negative sentiment is maintained (days) • The degree to which investors or authors respond negatively for stock investment	• If negative opinions occur continuously, the pessimistic prospects for the stock are maintained for a long time • Investors feel that the stock is unstable and seriously consider buying/selling • The occurrence probability of a credit event is relatively high
	The daily average range of fluctuation (x_3)		• The gap of sentiment value between opinions that occurred during one day • The difference of customers' response within short-term period (one day)	• The larger the range of daily fluctuation is, the more arguable investors are • The daily average shows variation in short-term
	The number of crossover (x_4)		• The number of sentiment crossover in a week • The changeover means that there are much of arguments	• A number of crossover means variation of overall opinions occurred within a week • Investors are recognizing issues related to risk event and arguing about it seriously
	The proportional gap between positive opinion and negative opinion (x_5)		• (The proportion of positive opinion) – (The proportion of negative opinion) • Overall tendency of opinions	• Overall tendency of investors • If the gap is positive, investors are showing optimistic view and there are favorable factor • However, in the case of negative gap, most investors expressed pessimistic view about the stock and guess there are some risk events
	The average growth rate of sentiment value by week	Positive sentiment (x_6)	• The variation of positive opinion in a week	• Relatively favorable factors or events has happened
		Negative sentiment (x_7)	• The variation of negative opinion in a week	• The rapidly growing aspect of negative opinions shows that serious credit event occurs

(dangerous, warning, and caution) based on the threshold value. According to each grade, the strategy for monitoring and responding when a credit event occurs will be different. Moreover, sentiment values that are contrary to each other in each database can be evidence of credit events. If opinions are opposite by the type of database, indicating that responses are different, then there is hidden information within a specific database, which can be a sign of the occurrence of credit events.

3) FORECASTING OCCURRENCE POSSIBILITY OF CREDIT EVENT

After identifying the sign of a credit event, the actual possibility of credit event occurrence is predicted by logistic regression. The prediction model based on logistic regression is composed of sub-indices for assessing risk at the prior step. First, data for different firms included in the same industry are collected to develop a prediction model through learning labeled data such as success or bankruptcy. In common with collecting data related to a target company, news and social data about other firms are collected, and then, the sentiment value of each indicator is calculated. Second, the regression equation for credit event occurrence is estimated through logistic regression. The dependent variable (y) is whether credit events such as bankruptcy and sudden organizational changes have occurred. The independent variables (x) are indicators (Table 2). Using the logistic regression equation, data of the target firm are put into the prediction model based on logistic regression. The probability of credit event occurrence is finally estimated. To validate the results derived from this prediction model, we developed a confusion matrix by comparing the actual number of incidences with the

TABLE 3. Case summary and the results of data collection.

Stock type			**KOSPI (Korean securities market)**		
Features			The capital owned by the business is more than $30 billion		
Core business			Shipping service composed of container segment and bulk segment		
Data	Period		2016.08.09–2017.08.08		
	Type of database		Before filtering		After filtering
	Objective DB	News article	370		283
	Subjective DB	Web community	914	81,950	60,362
		Bulletin board on portal site	80,370		
		SNS	666		
	Total		82,320		60,645

* No opinions were posted because this bulletin board is able to handle stocks listed on the exchange market except for stocks included in the over-the-counter market.

predicted number, which is higher than the cut-off probability and represented as 0 or 1 (binary).

IV. RESULTS

To illustrate this approach, the case of a firm was selected – Hyundai Merchant Marine – to cover the case of listed companies in the stock market. Hyundai Merchant Marine is a representative company for the marine transportation industry in South Korea. It is one of the companies in the shipping industry and is closely connected with Hanjin Shipping Co., Ltd., which went bankrupt in 2017. In addition, the global financial crisis caused risks in the shipping industry. Since uncertainties in the shipping industry can affect the variability of firms, the financial crisis could be on the rise.

A. DATA COLLECTION AND INFORMATION FILTERING

Data related to Hyundai Merchant Marine were collected from diverse databases, as shown in Table 3. To obtain objective data, news and numeric data, such as the stock prices and operating statuses of firms, were collected. In addition, subjective data, including those from SNSs and articles posted on portal sites, to reflect the opinions of general users, were collected through web scraping. Hyundai Merchant Marine has a large amount of data arising from a major crisis in the shipping industry (81,425 articles including both objective and subjective data). For raw data collected from the individual database, fake information with the possibility to instigate actions by individual investors or wrong information, such as advertisements, was eliminated by automatically finding abnormal periods and investigating their content. For this investigation, the abnormal period in which the number of opinions rapidly grew was extracted, and then, the contents during that period were analyzed in detail. While the news database was relatively clean and maintained objectively, social databases contained much fake information and noise. In particular, opinions posted on social databases are of short length, such as one sentence or phrase. If this sort of short information is accumulated, then clouded judgment occurs. Therefore, this short information was considered noise and then eliminated. A close look at the contents revealed common patterns in fake information – similar linguistic phrases and comments for specific authors who may convey hoaxes. The information used the same phrases but added some words, and they were written for the purpose of attempting to buy or sell for other investors.

As a result, in the case, there were three abnormal periods (shown in Figure 6). The first period occurred when a competitor of the firm filed for receivership. The second and third periods occurred because investors expected future profit-taking by encouraging other investors to buy or sell under the circumstances that the firm had repeated deficit operations and the competitor had no room for improvement despite various efforts from the government as well as the firm. During these periods, 21,292 short opinions were filtered and 296 fake information items (199 opinions at the 1st period, 48 opinions at the 2nd period, and 49 opinions at the 3rd period) were eliminated, while 87 news items, including advertisements and very short articles, were removed. Along with the case, seven irrelevant news articles were removed, and two opinions, an advertisement and a hashtag, which is a unique feature of SNS data, were eliminated.

B. RESULTS OF ANALYSIS

After filtering raw data, all documents were parsed and tagged according to natural language processing. The sentiment value [-1 (negative), 0 (neutral), $+1$ (positive)] was then assigned for each document in accordance with the fluctuation of the stock price (in the case of unlisted stocks, their selling price was used at this step). Each document's value was reassigned to each word using Naïve-Bayes rules, resulting in the sentiment value of each word. The value derived at this step is not a definitive value. This value is only used to select core keywords in pursuit of defining the sentiment value of each word considering relations between

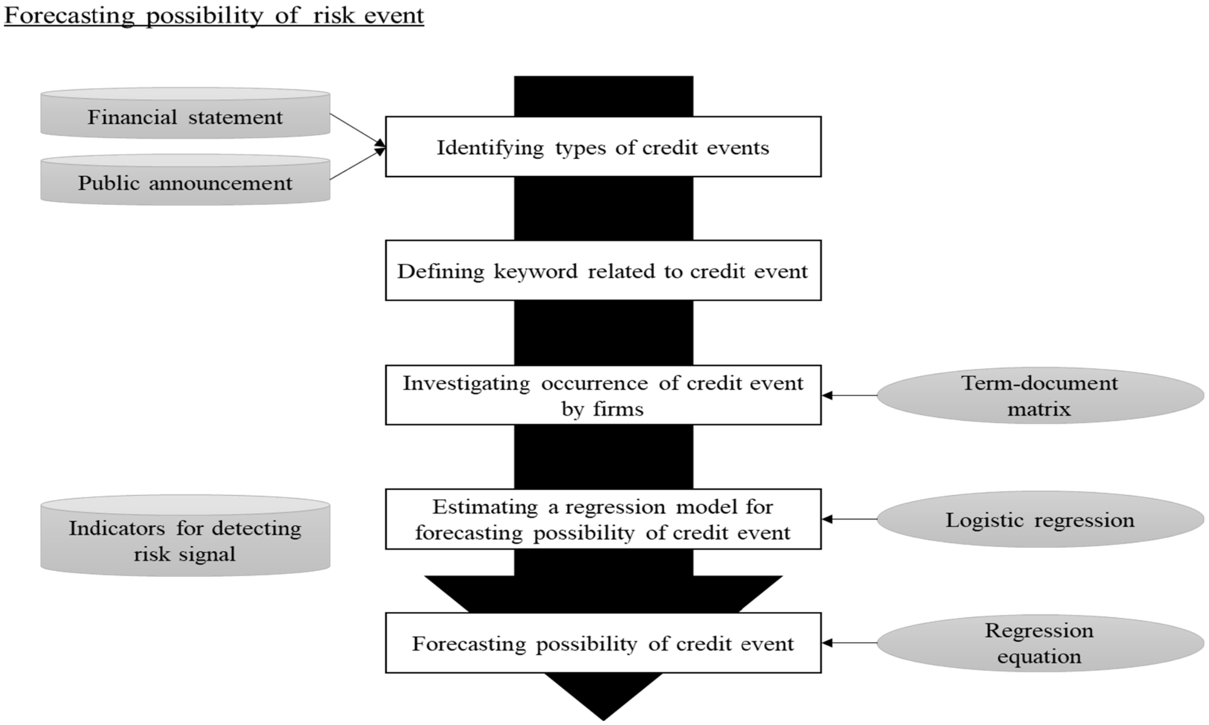
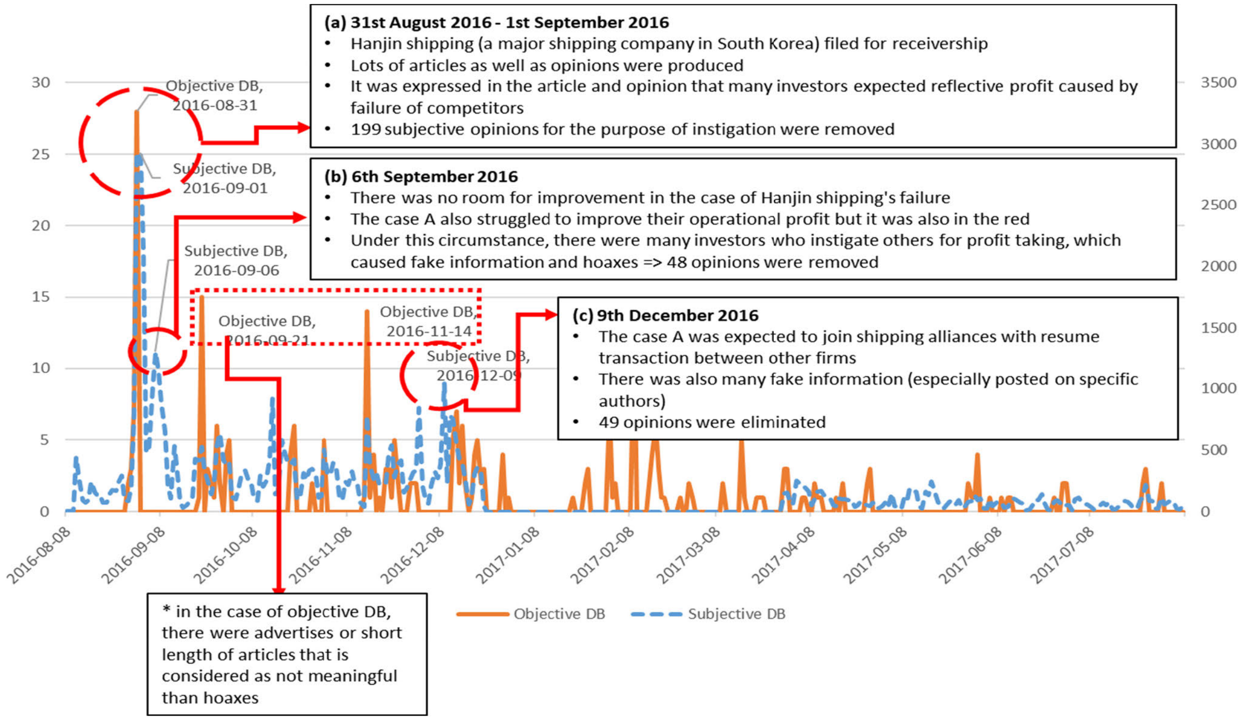

Forecasting possibility of risk event
Financial statement
Public announcement
Identifying types of credit events
Defining keyword related to credit event
Investigating occurrence of credit event by firms
Term-document matrix
Indicators for detecting risk signal
Estimating a regression model for forecasting possibility of credit event
Logistic regression
Forecasting possibility of credit event
Regression equation

FIGURE 5. Process for forecasting the possibility of credit events.



(a) 31st August 2016 - 1st September 2016
- Hanjin shipping (a major shipping company in South Korea) filed for receivership
- Lots of articles as well as opinions were produced
- It was expressed in the article and opinion that many investors expected reflective profit caused by failure of competitors
- 199 subjective opinions for the purpose of instigation were removed

(b) 6th September 2016
- There was no room for improvement in the case of Hanjin shipping's failure
- The case A also struggled to improve their operational profit but it was also in the red
- Under this circumstance, there were many investors who instigate others for profit taking, which caused fake information and hoaxes => 48 opinions were removed

(c) 9th December 2016
- The case A was expected to join shipping alliances with resume transaction between other firms
- There was also many fake information (especially posted on specific authors)
- 49 opinions were eliminated

* in the case of objective DB, there were advertises or short length of articles that is considered as not meaningful than hoaxes

FIGURE 6. Result of fake information filtering.

the words in sentences. If there is no stock price, the sentiment dictionary should be developed earlier. The sentiment value for each document and word can then be assigned based on the dictionary.

As a result, the core keywords with the highest sentiment value of Hyundai Merchant Marine, such as 'bond', 'Musk', 'terminal', and 'disposal', were derived. Due to the financial crisis brought about by Hanjin Shipping, their properties, such as terminals all over the world, were disposed of. Thus, many firms, including Hyundai Merchant Marine, competed to buy them. Hyundai Merchant Marine also tried to join a marine alliance with Musk. Thus, these firms are

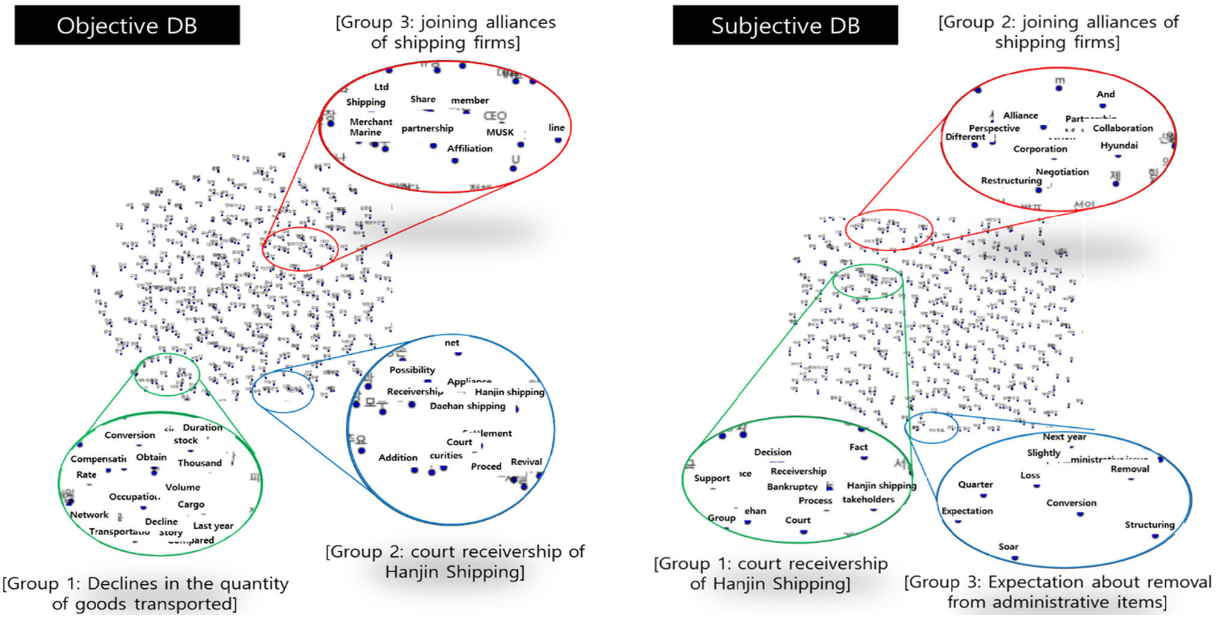


FIGURE 7. Result of word2vec.

concurrently mentioned in many documents. On the basis of preprocessed data, they were embedded by word2vec for each database - objective and subjective. As a result, a relation in the context between core keywords and other words was obtained. This relation is presented as the distance between the core keyword and each word. If a specific word is close to a core keyword, then it has a similar meaning or this word is written in the same context. The distance between words in word-embedding results becomes evidence for propagating the sentiment score of core keywords, which is allocated in reverse proportion to this distance. In the case of Hyundai Merchant Marine (in Figure 7), three word groups in the objective database – declines in the number of goods transported, joining an alliance of shipping firms, and the court receivership of Hanjin Shipping – were presented. The subjective database showed similar results. Expectations about the revocation of the administrative issue rose. Contrary to the situation with Hanjin Shipping, Hyundai Merchant Marine systems tried to improve profitability under legal supervision for a few months. Thus, this firm was expected to be unlisted from administrative issues from private investors. Depending on the word2vec results, the sentiment value of core keywords was propagated to each word based on the distance between core keywords and individual keywords. The value of each word was summed, and then, each document's sentiment was estimated. As a result of sentiment analysis, we found remarks such as rapid changes as time passed and conflicting sentiment by the database. As shown in Figure 8, the 1st and 2nd periods showed that objective DB had relatively positive opinions, while subjective DB contained negative reviews. These extraordinary phenomena can be evidence to detect abnormalities in financial and operational status.

In summary, seven indicators based on the number of opinions, keywords, and sentiment value were derived, as shown in Table 2. Using these indicators, logistic regression was conducted with data including five other firms in the shipping industry. Among them, four variables – days of similar sentiment (positive and negative) maintained, the number of sentiment crossovers, and the growth rate of positive opinions – were significant (Table 4). The logistic regression line is as follows (4):

$$\begin{aligned} &\text{Y (occurrence of credit event)} \\ &\quad = -1.791 - 0.369X_2 + 0.209X_3 - 0.372X_4 + 1.602X_6 \end{aligned} \tag{4}$$

This regression line made it possible to predict the occurrence of credit events by using events occurring in recent times. The predicted occurrence was then compared with the actual occurrence of the credit event. For example, on 21 October 2016, there was a public notice that Hyundai conducted capital reduction without refund, as shaded in Table 4. To predict the occurrence of this event, accumulated data and signals one month before and after this event were used. If the probability using a regression line depending on significant variables is higher than the threshold value, then it can be interpreted as the occurrence of credit events. Consequently, within two weeks, three similar credit events occurred. These events were also estimated by logistic regression, as shown in Table 4 and Figure 9, with an accuracy of 0.86667 and a precision of 0.75.

V. DISCUSSION AND IMPLICATION

A. STRATEGY ESTABLISHMENT

According to the predicted signal and risk grade, investors can establish strategies for investment. This study can help

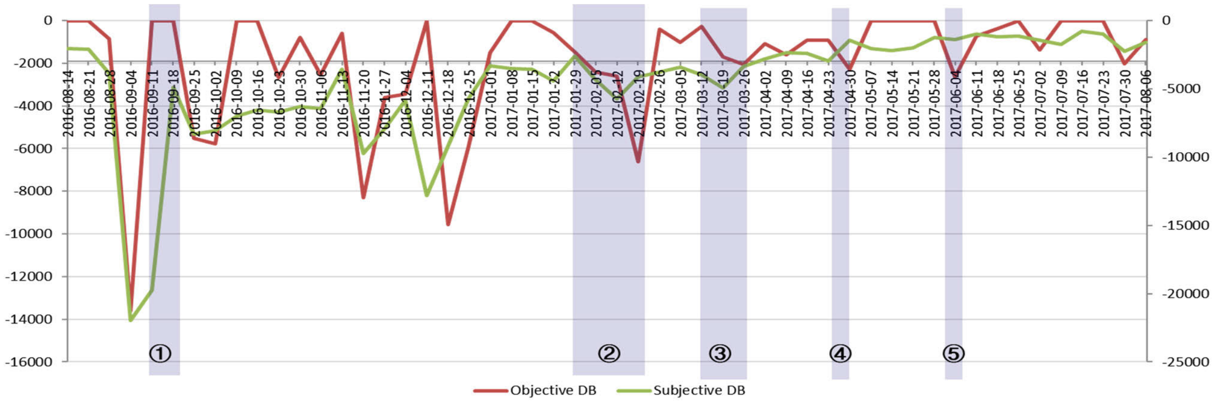


FIGURE 8. Trend of risk signal by the type of database (*Hyundai Merchant Marine*).

TABLE 4. Results of logistic regression in the case study.

time	X_2**	X_3**	X_4**	X_6**	logit	p	predicted	observed
signal in 2016-10-14	2	3	2	21%	-2.30958	0.090333	0	0
signal in 2016-10-15	2	3	2	7%	-2.53386	0.073518	0	0
signal in 2016-10-16	2	3	2	-1%	-2.66202	0.065252	0	0
signal in 2016-10-17	2	3	3	18%	-2.72964	0.061247	0	0
signal in 2016-10-18	2	4	2	13%	-2.22874	0.097199	1	1
signal in 2016-10-19	2	4	2	1%	-2.42098	0.081587	0	1
signal in 2016-10-20	2	5	2	-8%	-2.35616	0.086577	0	1
signal in 2016-10-21*	3	4	2	11%	-2.62978	0.067246	0	1
signal in 2016-10-22	3	4	2	-10%	-2.9662	0.048976	0	0
signal in 2016-10-23	3	3	2	-20%	-3.3354	0.034377	0	0
signal in 2016-10-24	3	4	1	-5%	-2.5141	0.074876	0	0
signal in 2016-10-25	2	3	2	-4%	-2.71008	0.062381	0	0
signal in 2016-10-26	1	4	3	2%	-2.40796	0.082568	0	0
signal in 2016-10-27	0	3	4	-5%	-2.7321	0.061106	0	0
signal in 2016-10-28	0	4	4	10%	-2.2828	0.092558	0	0

* The actual event (capital reduction without refund) occurred and was announced publicly.

** X_2: The period that positive sentiment is maintained, X_3: The period that negative sentiment is maintained, X_4: The number of sentiment crossovers, X_6: The average growth rate of positive sentiment value by week.

investors monitor and detect signals and then respond to credit events in advance to avoid expected losses or dominate profits. By using sub-indices, an integrated indicator for monitoring risks can be developed as shown below (5):

$$\text{Indicator for monitoring} = \sum_{i=1}^{n} W_i S_i \qquad (5)$$

After defining the weight for each signal metric, the weighted sum becomes the degree of a risk credit event. These weights can be determined in two ways: 1) subjective judgment and 2) financial metrics. The first relies on the opinions of investors. Investors can decide their weights differently with consideration of their importance. If investors focus on the growth rate of sentiment value, then they can give high weight to relevant metrics. The second way utilizes financial metrics with a discrimination index, which is presented in the form of an exponential function as follows (6):

$$\text{Weight considering financial metrics} = e^{1/(\mathit{financial\ metrics})} \qquad (6)$$

Financial metrics are a proximate indicator to identify the financial status of firms based on financial statements. These metrics serve as a weight for monitoring indicators based on opinion mining. There are diverse financial metrics, such as market capitalization, size or proportion of current assets, the rate of operating profits, debt ratio, and debt service coverage ratio. These metrics can be integrated by weighted sum through normalization and then converted to the form of weights. The higher this value is, the more stable the financial status of a firm is. The weight is derived by a reciprocal number of normalized weighted sums integrating each financial metric. The weight that reflects financial status makes the indicator for monitoring larger or smaller to observe more dramatic changes in signals. Like this trend, investors may catch risk signals and respond to them early.

The monitoring indicator with this weight suggests the degree to which the behavior pattern of investors and stability of management by firms serve as the criteria. By grading this indicator, the level of monitoring and strategy can be differentiated. If a specific firm has a high value for the monitoring indicator, especially if a signal is evaluated as 'dangerous', then intensive attention should be paid to this firm because many investors have shown quite negative opinions about this firm. Moreover, if various opinions – positive and negative – are continuously occurring, then a more intense and tight monitoring strategy is required for this stock. For that, the period of monitoring needs to be shortened, and the level of observation should be deeper for core risk events and surrounding keywords. On the other hand, in the case of a lower metric and showing a 'caution' signal, little precise attention is needed. Thus, a wait-and-see strategy might be more appropriate for this case because this sort of firm with a high value of monitoring indicator is stable in finance and operation, and investors' investment behavior is optimistic. While observing and monitoring opinions about this stock, the period for observation can be extended. However, for sudden and dramatic changes, it is necessary to examine events and how much sudden changes have been maintained since then.

B. EXTENSIBILITY TO DIFFERENT STOCK MARKETS

The proposed approach can be applied to stock markets in other places, including the United States, China, Japan, and Europe, if they have stock prices and opinions posted by investors. Since this approach mainly relies on customers' reviews, news, SNS posts, and reports about specific stocks, it is extendable to other markets. Although the information related to stock price for deciding sentiment value at the initial stage is not open publicly, it is possible to derive sentiments, in that opinion data can be utilized to establish a sentiment dictionary. This dictionary serves as baseline data for determining initial sentiment value before propagating the value of core keywords. For example, the stock market in the United States is also divided into national exchange markets such as the New York Stock Exchange (NYSE), National Association of Securities Dealers Automated Quotations (NASDAQ) market, OTC markets such as OTCBB (Over-The-Counter Bulletin Board), and Pink Sheets, similar to the South Korean stock exchange. For stocks exchanged in Pink Sheets, in common with Korean unlisted stocks, there is no obligation to reveal the financial status behind items that are traded, while the financial information for OTCBB items ought to be submitted and opened to investors. Although the information related to financial status is insufficient, it is possible to detect risk signals by relying on news and social data because there are many online communities and SNSs for sharing investment information (e.g., http://investing.com, http://www.etf.com, http://seekingalpha.com, https://finbiz.com).

In particular, the approach proposed in this study is not to predict an increase or decrease in stock price but to detect early signals for future risk event occurrence and help respond to such events appropriately. Although stock prices and financial metrics are very important, it is more important to identify investors' behavioral patterns and trends from the perspective of ex-ante prediction and planning. Thus, information based on news and opinions posted by investors ranging from individual investors to experts could be enough to catch risk signals in advance.

VI. CONCLUSION

In summary, this paper suggests a new algorithm to support decision-making in stock investment by detecting early signals and predicting the occurrence probability of credit events through opinion mining and logistic regression models. For a long time, news and official reports published by securities analysts or stock experts have been important and ample sources when investors decide to buy or sell stocks. However, with the rise of IT devices and growth in SNS use, information or individual intention is vigorously shared through online communities and private SNSs when making decisions about stock investment. In particular, some bad investors can deliberately write and deliver incorrect information for the purpose of manipulating the stock price. Such a capitulation bottom may cause incorrect investment and hurt individual investors. Thus, it is critical to filter out fake information and catch accurate signals in advance. Consequently, we proposed an algorithm for detecting risk signals early based on a large amount of opinion data.

From the viewpoint of data, our approach contributes to extending the availability of subjective data such as that of SNSs and online communities when making an investment in stocks. Social data have been widely used to discover customers' satisfaction with or complaints about products or services, although the applicability of social data is limited to customer management, quality management, or new product/service development. In particular, social data generated from SNSs are narrowly applied to analyze trends in stock investment in previous studies, although most investors can take advantage of diverse information posted on websites or online communities when deciding to buy or sell stocks. At this point, we utilized user-generated data on social media that are valuable for identifying human behavior and preference through web communities related to stock investment. This type of database provides increasingly valuable information by supporting the recognition of different human behaviors in the stock market, such as instigation. Thus, our paper expanded the availability of social data in the finance market by reflecting a mechanism for when investors make decisions about whether to trade stocks. Moreover, the availability of social data enables us to analyze and obtain insights for private stocks traded on the over-the-counter market, where accessibility about information is quite limited.

From the viewpoint of data processing, our paper addresses fake information because the stock market is influenced by information between investors and there is much information regardless of honesty. It is necessary to cope with the credibility of the data itself, which is considered fake information filtering in this paper. This process is not just screening unnecessary or irrelevant information but also filtering real false information that may have an impact on decision-making for stock investment. Graph-based semi-supervised learning supported the classification of sentiment of words or documents more accurately in a situation, especially the word2vec-based learning context in the stock market. In other words, specific words and relations between words might have different meanings in other situations. However, graph-based semi-supervised learning helps to learn and classify words or documents situationally. Words with a high sentiment value served as labeled data, and their values were propagated to other close words with no label that were regarded as having ambiguous meanings. Graph-based semi-supervised learning makes other data locally smooth and consistent with the labeled data. The logistic regression-based prediction model using historic finance events, such as bankruptcy and the status of board members, enables the identification of relations between financial events and success and failure. This model showed which events affect stock investment and furthermore the most influential events that contribute to predicting the success or failure of each stock. The prediction model in this paper supported the derivation of statistically significant prediction results within the framework of the stock market. The integrated model in this paper increases the likelihood of applicability and availability in a similar context.

Although most subjective opinions may cause biased judgment, comprehensive analysis of social data is able to provide practical insights and signals that need tight monitoring. This is because a bulk of subjective opinions can show the direction of the investment behavior of authors (or investors) and encourage other investors to buy or sell their stocks by providing subjective insights and intelligence, such as collective intelligence. Thus, our approach can be regarded as a useful tool for decision-making in stock investment based on behavioral patterns and collective intelligence.

Moreover, this algorithm will be helpful for detecting hidden risk without financial status. Although each firm must provide information about its financial condition and important changes in operations and management, some firms may hide their unfavorable status. Thus, some events are not exposed outwardly or are hidden, making it difficult to detect these hidden events, which may cause a serious loss. For these situations, the financial statement is not appropriate. It can be regarded as an outdated source for predicting future risk and preventing great loss. To ex-ante predict and discover early signals, another resource is required. The social database provides such a source.

This paper contributes to detecting signals in advance based on social data. Because the stock market is sensitive to investors' behavior as well as firms' status, social data serve as an important proxy for the behaviors of investors. Most people currently represent their opinions on social media, and social data are a critical database to identify human behavior and preference. Because previous studies have focused on stock prices and official news, it is impossible to detect signals without financial information. However, this paper depends on social data, and predicting in the over-the-counter market is possible using only social data. In this paper, the proposed algorithm focuses on the stock market, but it can be applied to other environments that are largely influenced by human behavior, such as social media commerce. Although previous studies have focused on predicting the rise and decline of stock prices, our study supports making decisions for stock investment. The proposed algorithm can be applied to both personal buyers and firms, which makes it possible for the government to deal with financial crises at the national level.

Nevertheless, much more research needs to be done. First, the process for filtering fake information should be further elaborated in different ways. The proposed approach narrowly focused on the periods in which opinions were rapidly growing. Thus, it was difficult to find and screen unnecessary or fake information in other periods. Because of this, the remaining but wrong information may lead to bad decisions on whether to buy or sell stocks. Thus, more effort is needed to purify a large amount of raw data that has accumulated for a long time. Second, little attention was paid to establishing an investment strategy because this paper intended to develop an algorithm for detecting early signals and supporting the decision-making process. While the proposed algorithm can provide baseline information such as risk or opportunity for stock investment early, the detailed strategy is unavailable. In particular, suggested strategies were related to whether to buy or sell. However, the risk for specific stocks, the period for monitoring, and the sort of risk need to be handled. A complementary practical strategy with assistance from a financial analyst or other experts is needed to align practical strategy and goals. Third, the sentiment dictionary needs to be refined for each industry by using information from a wide range of databases. As mentioned above, there is an ambiguity of words according to the type of industry because each industry has distinct characteristics. Although the sentiment dictionary in this paper was based on a database related to a specific industry, it should be further elaborated and refined by utilizing another database for generalization in the near future.

REFERENCES

[1] H. Löbler, "When trust makes it worse—Rating agencies as disembedded service systems in the U.S. financial crisis," *Service Sci.*, vol. 6, no. 2, pp. 94–105, Jun. 2014.

[2] W. N. Moulton and H. Thomas, "Bankruptcy as a deliberate strategy: Theoretical considerations and empirical evidence," *Strategic Manage. J.*, vol. 14, no. 2, pp. 125–135, Feb. 1993.

[3] G. A. A. J. Alkubaisi, S. S. Kamaruddin, and H. Husni, "Conceptual framework for stock market classification model using sentiment analysis on Twitter based on hybrid Naïve Bayes classifiers," *Int. J. Eng. Technol.*, vol. 7, no. 2.14, pp. 57–61, 2018.

[4] G. A. Alkubaisi, S. S. Kamaruddin, and H. Husni, "Stock market classification model using sentiment analysis on Twitter based on hybrid naive Bayes cLASSIFIERS," *Comput. Inf. Sci.*, vol. 11, no. 1, pp. 52–64, 2018.

[5] A. Mittal and A. Goel. (2011). Stock Prediction Using Twitter Sentiment Analysis. Standford University, CS229. Accessed: 2012. [Online]. Available: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[6] T. Oberlechner and S. Hocking, "Information sources, news, and rumors in financial markets: Insights into the foreign exchange market," *J. Econ. Psychol.*, vol. 25, no. 3, pp. 407–424, Jun. 2004.

[7] C. M. Daily, "Governance patterns in bankruptcy reorganizations," *Strategic Manage. J.*, vol. 17, no. 5, pp. 355–375, May 1996.

[8] R. J. Taffler and M. Tseung, "The audit going concern qualification in practice-exploding some myths," in *Proc. Accountants' Mag.*, 1984, pp. 263–269.

[9] K. Raghunandan and D. V. Rama, "Audit reports for companies in financial distress: Before and after SAS No. 59," *Auditing*, vol. 14, no. 1, p. 50, 1995.

[10] T. E. Mckee, "Developing a bankruptcy prediction model via rough sets theory," *Int. J. Intell. Syst. Accounting, Finance Manage.*, vol. 9, no. 3, pp. 159–173, 2000.

[11] C. M. Daily and D. R. Dalton, "Corporate governance and the bankrupt firm: An empirical assessment," *Strategic Manage. J.*, vol. 15, no. 8, pp. 643–654, Oct. 1994.

[12] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Syst. Appl.*, vol. 135, pp. 60–70, Nov. 2019.

[13] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decis. Support Syst.*, vol. 55, no. 3, pp. 685–697, Jun. 2013.

[14] S. L. Heston and N. R. Sinha, "News vs. Sentiment: Predicting stock returns from news stories," *Financial Analysts J.*, vol. 73, no. 3, pp. 67–83, Jul. 2017.

[15] S. Feuerriegel and J. Gordon, "Long-term stock index forecasting based on text mining of regulatory disclosures," *Decis. Support Syst.*, vol. 112, pp. 88–97, Aug. 2018.

[16] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, Feb. 2009.

[17] E. Junqué de Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," *Inf. Process. Manage.*, vol. 50, no. 2, pp. 426–441, Mar. 2014.

[18] R. Bose, A. Das, J. Poray, and S. Bhattacharya, "Risk analysis for long-term stock market trend prediction," in *Proc. Int. Conf. Adv. Comput. Data Sci.*, Ghaziabad, India: Springer, 2019, pp. 381–391.

[19] L. Liu, J. Wu, P. Li, and Q. Li, "A social-media-based approach to predicting stock comovement," *Expert Syst. Appl.*, vol. 42, no. 8, pp. 3893–3901, May 2015.

[20] Y. Qian, X. Deng, Q. Ye, B. Ma, and H. Yuan, "On detecting business event from the headlines and leads of massive online news articles," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102086.

[21] S. Han, X. Hao, and H. Huang, "An event-extraction approach for business analysis from online Chinese news," *Electron. Commerce Res. Appl.*, vol. 28, pp. 244–260, Mar. 2018.

[22] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proc. 12th Int. Conf. World Wide Web*, 2003, pp. 519–528.

[23] V. L. Miguéis and H. Nóvoa, "Exploring online travel reviews using data analytics: An exploratory study," *Service Sci.*, vol. 9, no. 4, pp. 315–323, Dec. 2017.

[24] X. Yang, Y. Zhu, and T. Y. Cheng, "How the individual investors took on big data: The effect of panic from the Internet stock message boards on stock price crash," *Pacific-Basin Finance J.*, vol. 59, Feb. 2020, Art. no. 101245.

[25] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[26] H. Chen and D. Zimbra, "AI and opinion mining," *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 74–80, May 2010.

[27] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

[28] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2016, pp. 452–455.

[29] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020.

[30] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[31] C. Lim and P. P. Maglio, "Data-driven understanding of smart service systems through text mining," *Service Sci.*, vol. 10, no. 2, pp. 154–180, Jun. 2018.

[32] A. Esuli and F. Sebastiani, "SentiWordNet: A high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, no. 1, p. 26, 2007.

[33] X. Ding, B. Liu, and L. Zhang, "Entity discovery and assignment for opinion mining applications," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1125–1134.

[34] B. Zhang, Y. Bai, Q. Zhang, J. Lian, and M. Li, "An opinion-leader mining method in social networks with a phased-clustering perspective," *IEEE Access*, vol. 8, pp. 31539–31550, 2020.

[35] H. H. Binali, C. Wu, and V. Potdar, "A new significant area: Emotion detection in E-learning using opinion mining techniques," in *Proc. 3rd IEEE Int. Conf. Digit. Ecosystems Technol.*, Jun. 2009, pp. 259–264.

[36] D. Lee, O.-R. Jeong, and S.-G. Lee, "Opinion mining of customer feedback data on the Web," in *Proc. 2nd Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2008, pp. 230–235.

[37] E. Yu, Y. Kim, N. Kim, and S. R. Jeong, "Predicting the direction of the stock index by using a domain-specific sentiment dictionary," *J. Intell. Inf. Syst.*, vol. 19, no. 1, pp. 95–110, Mar. 2013.

[38] A. Zubiaga and A. Jiang, "Learning class-specific word representations for early detection of hoaxes in social media," 2018, *arXiv:1801.07311*. [Online]. Available: http://arxiv.org/abs/1801.07311

[39] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[40] V. L. Rubin and N. Conroy, "Discerning truth from deception: Human judgments and automation efforts," *1st Monday*, vol. 17, no. 5, pp. 1–23, 2012.

[41] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, vol. 2, 2012, pp. 171–175.

[42] L. Zhou and D. Zhang, "Following linguistic footprints: Automatic deception detection in online communication," *Commun. ACM*, vol. 51, no. 9, pp. 119–122, Sep. 2008.

[43] C.-H. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4321–4330, Apr. 2009.

[44] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.

[45] M. Yang, Z. Lu, X. Chen, and F. Xu, "Detecting review spammer groups," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 5011–5012.

[46] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 1242–1247.

[47] I. Lašek and P. Vojtáš, "Semantic information filtering-beyond collaborative filtering," in *Proc. 4th Int. Semantic Search Workshop*, 2011, pp. 1–8.

[48] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1234–1244, Jul. 2019.

[49] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," 2017, *arXiv:1704.07506*. [Online]. Available: http://arxiv.org/abs/1704.07506

[50] M. Bosma, E. Meij, and W. Weerkamp, "A framework for unsupervised spam detection in social networking sites," in *Proc. Eur. Conf. Inf. Retr.* Barcelona, Spain: Springer, 2012, pp. 364–375.

[51] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on Twitter," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 153–164.

[52] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama, ''Assessment of tweet credibility with LDA features,'' in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 953–958.

[53] Y. Chen, Y. Zhao, B. Qin, and T. Liu, ''Product aspect clustering by incorporating background knowledge for opinion mining,'' *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0159901.

[54] X. Zhu and A. B. Goldberg, ''Introduction to semi-supervised learning,'' *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.

[55] O. Chapelle, B. Schölkopf, and A. Zien, ''A discussion of semi-supervised learning and transduction,'' in *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006, pp. 473–478.

[56] A. Hussain and E. Cambria, ''Semi-supervised learning for big social data analysis,'' *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018.

[57] D. Yu, N. Chen, F. Jiang, B. Fu, and A. Qin, ''Constrained NMF-based semi-supervised learning for social media spammer detection,'' *Knowl.-Based Syst.*, vol. 125, pp. 64–73, Jun. 2017.

[58] J. K. Rout, A. Dalmia, K.-K.-R. Choo, S. Bakshi, and S. K. Jena, ''Revisiting semi-supervised learning for online deceptive review detection,'' *IEEE Access*, vol. 5, pp. 1319–1327, 2017.

[59] X. Zhu, ''Semi-supervised learning with graphs,'' Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2005.

[60] P. D. Chowdary, D. L. Che, and B. Cui, ''Neurotrophin signaling via long-distance axonal transport,'' *Annu. Rev. Phys. Chem.*, vol. 63, no. 1, pp. 571–594, May 2012.

BYUNGUN YOON (Senior Member, IEEE) is currently a Professor with the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea. His research interests include patent analysis, new technology development methodology, and visualization algorithms.

YUJIN JEONG is currently a Postdoctoral Researcher with the Department of Industrial and Systems Engineering, Dongguk University. Her research interests include SMEs, quality management, open innovation technology forecasting, technology intelligence, data mining, and patent analysis.

SUNHYE KIM is currently pursuing the Ph.D. degree with the Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea. Her research interests include patent analysis, text mining, technology intelligence, and natural language processing.