Asia Pacific Journal of Information Systems
Vol. 32 No. 4 (December 2022), 770-802

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
https://doi.org/10.14329/apjis.2022.32.4.770

Applications of the Text Mining Approach to Online Financial Information

Hansol Lee[a], Juyoung Kang[b,*], Sangun Park[c]

[a] *Assistant Professor, School of Business, Ajou University, Korea*
[b] *Professor, School of Business, Ajou University, Korea*
[c] *Professor, Department of MIS, Kyonggi University, Korea*

ABSTRACT

With the development of deep learning techniques, text mining is producing breakthrough performance improvements, promising future applications, and practical use cases across many fields. Likewise, even though several attempts have been made in the field of financial information, few cases apply the current technological trends. Recently, companies and government agencies have attempted to conduct research and apply text mining in the field of financial information. First, in this study, we investigate various works using text mining to show what studies have been conducted in the financial sector. Second, to broaden the view of financial application, we provide a description of several text mining techniques that can be used in the field of financial information and summarize various paradigms in which these technologies can be applied. Third, we also provide practical cases for applying the latest text mining techniques in the field of financial information to provide more tangible guidance for those who will use text mining techniques in finance. Lastly, we propose potential future research topics in the field of financial information and present the research methods and utilization plans. This study can motivate researchers studying financial issues to use text mining techniques to gain new insights and improve their work from the rich information hidden in text data.

Keywords: Text Mining, Finance, Fiscal Policy, Financial Information, Keyword Analysis, Social Network Analysis, Sentiment Analysis, Topic Modeling, Machine Learning

Ⅰ. Introduction

1.1. Research Background

With the development of information and communication technology, many forms of data that qualify as big data are currently being created. Because big data contain a wealth of information, research, and projects based on analyzing these data are being conducted. In accordance with this trend, several studies are using big data in the financial services industry. For example, in recent studies of stock price pre-

*Corresponding Author. E-mail: jykang@ajou.ac.kr

dictions, huge amounts of data reflecting political and social trends, which affect stock markets, are analyzed to predict stock price fluctuations (Schumaker et al., 2012).

Big data include not only numerical data but also unstructured data that are difficult to analyze. Unstructured data include materials such as text, pictures, and videos. Recent studies draw large amounts of information and important implications from unstructured data that cannot be derived from other available data. Text mining has become necessary because most of the content produced by various reports and news is expressed as data that cannot be simplified to standardized numbers and because posts on blogs or social network services (SNSs), which provide large quantities of information about public opinions, are also composed of unstructured text data.

Essentially, the range of data that can be used for analysis is expanding to include qualitative data, with greater interest in text data. A recent study estimated that about 80% of the explosive growth in data is driven primarily by text-based data (Guo et al., 2014). The news articles, blogs, community posts, social network posts, and reports that are commonly encountered are all textual data. Unlike numerical data, textual data contain a wide variety of information. Because the text is composed of natural language, which is difficult to analyze, only a few studies of textual data have been conducted(Hearst, 1999; Song, 2017). Owing to improvements in computing power and the introduction of machine learning techniques, however, it has become common to use text mining in many fields, including finance. It is important for researchers to study in the financial context because only a small fraction of the application of text mining to financial studies is discussed in finance journals (Aziz et al., 2022).

According to Rutgers Business School, big data, including internal or external data that are not summarized, can be utilized in accounting studies. Additionally, keyword and social network analyses may also be useful for such studies (Rugters, 2015). Nassirtoussi et al. (2014) studied market predictions using textual data from the Internet and showed that various forms of textual data, such as tweets, news articles, financial reports, and blog posts, can be used in financial studies.

In recent years, natural language processing techniques, text summarization methods, and quantitative evaluations of financial texts, such as readability and formality checks, have been introduced to create additional value. In addition, investigations of dynamic textual data that change over time provide richer and more accurate information for decision-making. For instance, Oracle Financial Services Enterprise Fraud Management uses various technologies, such as artificial intelligence, machine learning, and natural language processing to evaluate and analyze real-time textual data (Oracle, 2022).

1.2. Research Objectives

Text mining facilitates effective and efficient analysis of textual data. Because textual data consists of unstructured information such as news articles, online posts, and words in video or audio, many researchers have attempted to use text mining techniques in various areas of research, including financial research. However, despite the fact that text data has a wealth of information not found in general structured data, it is difficult to extract and use such information, so financial studies analyzing structured data such as financial statements are still more common.

In this study, we review studies to determine the use of text mining in the financial sector, investigate the main methodologies, and describe each practice

using related works. We also provide practical examples evaluating South Korea's fiscal policy to demonstrate the various text mining technologies with an outlook on how the results can be applied. Finally, we suggest academic and practical directions for future research in finance using text mining.

The rest of this paper is organized as follows: In Section 2, we examine the characteristics and procedures of text mining and investigate several studies affiliated with text mining in finance; Section 3 presents various text mining technologies applicable in the field of financial information; Section 4 illustrates empirical cases applying text mining to evaluate government policies related to financial execution. Section 5 proposes multiple topics for future financial research using text mining. Finally, in Section 6, we discuss the study contributions and recommend directions for future studies.

Ⅱ. Literature Review

2.1. Characteristics of Textual Data and Text Mining

Text mining is a methodology that derives valuable information or patterns for decision-making through text data, such as emails, customer reviews or feedback, research reports, and SNS posts. Because this methodology is based on computational linguistics, statistics, big data analytics, and machine learning techniques, the desired information can be derived accurately and in detail (Miner et al., 2012).

In recent years, the volume of available textual data has skyrocketed with the increasing popularity of online communities, such as SNSs. This large amount of textual data is valuable for analysis as it contains a wealth of information. Identifying keywords and critical topics in various documents by hand is time-consuming and expensive. In the case of web data, which differs from well-organized reports or research papers, identifying key information using only manpower is difficult. However, text mining can save time and money in deriving the main topics from vast quantities of documents or classifying documents that contain similar characteristics.

In addition, text mining helps to derive potentially valuable information that was previously difficult to obtain. For example, topic modeling, a text mining technique, can derive major issues mentioned in a large amount of text, including those that are already known and those that are important but difficult to identify by hand. Thus, text mining allows a researcher to detect both frequently and infrequently voiced public opinions in a large amount of text.

Sentiment analysis can categorize the polarity of a given text, that is, whether the underlying sentiment of the text is positive or negative. Sentiment analysis can be used as a key indicator for evaluating national fiscal execution based on online content related to government policy. In addition, by analyzing customers' experiences, companies can determine customers' emotions regarding products or services.

Some of the information available through text mining was previously obtained through controlled experiments or interviews. Surveys, telephone interviews, and experiments yield results based on the structure designed by the researcher. However, the data selected for text mining can include newspaper articles, research data, reports, meeting minutes, or SNS posts, which are created for their own purposes. According to Connelly et al. (2016), the data obtained for big data analyses are "found data" and not the same as data created by researchers. Moreover, the biases of researchers and respondents are not present in these data. These features of big data can solve

many of the problems encountered in traditional surveys or experiments.

For example, social desirability is one of the main biases that distort survey results. Voters prejudiced against Barack Obama and Hillary Clinton in the 2008 U.S. presidential primary elections were unlikely to voice their prejudices against African‐Americans or women. Stephens-Davidowitz (2014) and Stephens-Davidowitz (2017) explain this phenomenon using Google search results to demonstrate the distortion of survey results due to social desirability bias. In a preliminary survey of voters forecasting the 2016 U.S. presidential election, the results showed that few voters supported Donald Trump, and his chances of being elected were extremely low as compared to those of Hillary Clinton (FiveThirtyEight, 2016; Silver, 2016). However, according to Google search results, several texts agreed with Donald Trump, and Seth Meyers (an American television host) predicted Donald Trump's victory. As a result, analyses of online textual data drew more accurate conclusions regarding the 2016 U.S. presidential election than surveys based on more established methods. Thus, by reducing the biases of researchers and respondents, text mining is useful for diagnosing and evaluating current situations.

2.2. Text Mining Research Procedure

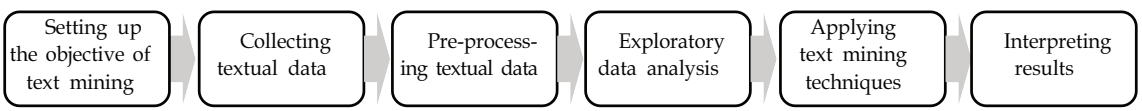
Text mining research starts with clearly establishing the goals of the analysis. Unclear analysis goals render the text mining process inefficient. In addition, it is necessary to check whether the problem can be solved using text mining techniques. This step is closely related to the next step, data collection, as it may be difficult to collect textual data about the object to be analyzed. Some problems cannot be analyzed because the amount of data is too small. Thus, the most important aspect of text mining research is selecting an object for analysis and a goal.

After setting the research objective and collecting textual data, it is necessary to process the text, which is in the form of a human language, into a form that can be understood by a computer. This step is called "data preprocessing." Because text mining requires using statistics and algebraic techniques, the results are very likely to be meaningless if rigorous data preprocessing is neglected.

Next, exploratory data analysis can be conducted to derive information from the preprocessed data. Furthermore, additional information can be obtained by performing suitable procedures for the analysis, such as classification and clustering. Importantly, in text mining, the interpretation of results requires careful attention, as the researcher's bias can lead to incorrect conclusions during data purification and analysis. Thus, research must be conducted according to the original study goals, but the results must be interpreted based on the evidence presented without contradicting the core tenet of research ethics. <Figure 1> shows the overall procedure for text mining research.

2.3. Text Mining Applications in Finance

Turner et al. (2013) found that, as of 2010, only banks and financial institutions used big data analysis



<Figure 1> Procedure for Text Mining Research

to enhance their competitiveness; however, by 2012, about 71% of institutions had adopted big data analysis, implying a growth rate of 97% compared to 2010. In recent years, increasingly many organizations have used big data analysis as an essential competence. Specifically, in the financial sector, big data analysis techniques are widely used in various capacities, such as social media analysis, web analysis, risk management, fraud detection, and intelligent security systems. Special information that was previously unknown is expected to be obtained through text mining (Pejic-Bach et al., 2019).

Because text mining analyzes text, a broad and diverse range of data can be used. In addition to information from various studies and reports, new information can be derived from text data in natural languages, such as news articles, customer opinions, and expert interviews. Several recent studies have specifically analyzed social media. Social media text data are meaningful because, unlike well-refined news articles and reports, they comprise opinions from the general public. Previously, investigating public opinion required conducting a large-scale questionnaire; however, such studies take a long time and are expensive because developing a questionnaire survey requires prior research and preparation. However, analyzing an SNS, in which many thoughts and opinions are recorded, allows a more efficient survey of public opinion. In addition, the scope of data for text analysis, including audio data, interviews, and public opinion surveys, is not limited to textual data and is gradually expanding.

2.3.1. Text Mining Research in Finance

Text mining techniques are widely used in various fields, and the use of text mining in financial studies is growing exponentially. According to Pejić Bach et al. (2019), prior to 2014, text mining research in the financial sector was limited; however, it has been steadily increasing since 2015, and the citation index of financial research using text mining has continued to increase as well.

Financial research using text mining can be classified by methodology. Text mining techniques include keyword, sentiment, topic, and social network analyses. Keyword analysis is a fundamental and important technique in text mining and encompasses mainly two different types of approaches: (i) descriptive statistical approaches based on word frequency or term frequency-inverse document frequency (TF-IDF); (ii) linguistic approaches, such as n-gram techniques or co-occurrences of words. In addition, these approaches can be used together. Social network analysis is the process of drawing key information from the relationships among words, topics, and authors in a set of textual data. Centrality analysis is often used to identify core information, such as keywords, the main idea, and the influencer in a network.

Alongside this, machine learning techniques are widely used in text mining technologies such as embedding, classification, and clustering. In fact, these are baseline techniques that derive the results of sentiment or topic analysis. The novel text mining approaches with machine learning techniques allow the expansion of research areas in the finance sector. Sentiment analysis is a method for estimating the attitude or emotion of a group of texts or the sentiment of texts' authors by analyzing the texts. This method is also called opinion mining (Pang and Lee, 2009). It mainly classifies the emotions in a text as positive and negative and has the advantage of being able to predict or evaluate social phenomena through a ratio. Topic analysis is deriving the topics contained in a text through probabilistic inference. In the past, human beings directly read and categorized texts to

<Table 1> Use of Text Mining in Financial Studies

Methodology	Author	Objectives
Keyword Analysis	Mihalyi and Mate (2019)	This study analyzed the International Monetary Fund's National Report (Article IV), comparing the existing research method with a method using word frequency. It showed that future institutional changes can be predicted by analyzing changes in the International Monetary Fund's policy recommendations over time.
	Shirata et al. (2011)	This study attempted to predict corporate bankruptcy by analyzing textual data rather than financial information. The study predicted bankruptcy by applying word frequency analysis and natural language processing techniques to Japan's annual financial reports.
	Junqué de Fortuny et al. (2014)	This study compared the models for stock price prediction, and instant state-of-the-art text mining technologies were introduced. All the articles published online in major Flemish newspapers from 2007 to March 2012 were analyzed. The bag-of-words model and document sentiment polarity were used to analyze the articles. It was found that text mining and technical indicators can be used to oversee stock price movement directionality.
Social Network Analysis	Rönnqvist and Sarlin (2015)	This study analyzed the interdependence of banks from the perspective of bank connections or risk using text-based network analysis, whereas previous studies of this topic used numerical data. The study performed a quantitative analysis based on the names of banks mentioned at the same time in documents, a visualization, and a centrality analysis to derive the importance of banks and changes in their importance over time.
	Mao et al. (2015)	This study measured the mutual influence of bank customers through social network analysis. In addition, it conducted a centrality analysis to analyze bank customers' influences on each other to identify the most influential customers. The study revealed, through network analysis, that customers have direct and close relationships with each other but that a small number of customers form an overall relationship network.
Sentiment Analysis	Davis et al. (2012)	This study counted the number of pessimistic and optimistic words by analyzing the major press releases of about 23,000 companies by quarter. This analysis showed that the frequency of optimistic words could predict corporate performance in the next quarter.
	Tetlock (2007)	This study introduced sentiment analysis to measure the relationship between news media and the stock market. It found that pressure on stock prices decreases as more pessimistic expressions in the media increase and that the frequency of pessimistic expressions that are abnormally high or low is helpful in predicting market volume.
	Chen et al. (2018)	This study analyzed posts on Twitter and Weibo to derive topics related to the stock market. It derived these topics by searching for meanings from various media data generated by general users. The study effectively identified policy changes in China that significantly impacted the Shanghai Stock Trading Composite Index.
Topic Modeling	Moro et al. (2015)	This study used topic modeling techniques to derive trends in the use of business intelligence in the banking industry. The study obtained groups by topic from 219 studies registered between 2002 and 2013 and suggested major trends that can be derived from this large set of documents while minimizing manual work. This research occupies a dominant position and shows that there is great interest in minimizing risk through the detection or prevention of fraud and bankruptcy.
	Nguyen and Shirai (2015)	This study attempted to predict the stock market by applying topic modeling based on sentiment analysis to social media. It used the joint sentiment/topic model and the topic sentiment latent Dirichlet allocation model, which are emotion-topic models designed to derive topics for each emotion.

<Table 1> Use of Text Mining in Financial Studies (Cont.)

Methodology	Author	Objectives
	Aziz et al. (2022)	This study successfully clustered topics from the relevant articles and constructed the evolution of topics over time using topic modeling. It provides various research topics regarding financial studies that use machine learning techniques.
Machine Learning Approach	Nyman et al. (2021)	In this study, the authors studied large amounts of textual data to estimate the impacts of narratives and sentiment on developments in the financial system. This study shows that text mining would be useful in evaluating risks to financial stability.
	Fu et al. (2020)	This study predicted and identified potential default risk platforms. Keywords were extracted from investor comments using deep learning neural networks. The study used a bidirectional long/short-term memory model to make accurate default risk predictions of platforms.
	Qian et al. (2019)	In this study, a model was built to detect business events based on the clustering-annotation-classification strategy. It classified potential events from news headlines using information about business events that were extracted by exporting annotations on a group of terms derived from the result of clustering.
	Xu et al. (2020)	This study introduced a new framework for sentiment classification based on a continuous naïve Bayes learning technique. It analyzed large-scale and multi-domain e-commerce platform product reviews.

determine the topics that they contained. Thus, the longer a text is, the more inefficient this process is, and human bias may cause issues in some cases. However, through probabilistic inference, it is possible to derive the main topics in a text and those omitted owing to human bias or ignorance. <Table 1> shows the uses of text mining in the financial sector by methodology.

From <Table 1>, the studies indicate that the text mining approach can be utilized to obtain new insights and valuable information compared to the common structured data. Mihalyi and Mate (2019) gained significant insights by using simple text analysis techniques. Shirata et al. (2011) and Junqué de Fortuny et al. (2014) were able to predict the market using text mining rather than using the general approach with structured data.

One of the advantages of the text mining technique is being able to build networks using important keywords. Rönnqvist and Sarlin (2015) and Mao et al. (2015) drew networks using lexical collocations and found hidden connections in the financial market. Text data can also be applied to machine learning techniques such as classification, prediction. Davis et al. (2012), Tetlock (2007), and Chen et al. (2018) focused on the relationship between online responses and the market by classifying the keywords into positive and negative response and evaluating the market based on the emotion.

In addition, the new financial information can be derived by the statistical approaches such as topic modeling and machine learning. Nguyen and Shirai (2015) used topic modeling approach to improve the accuracy of stock price predictions, suggesting that greater predictive power can be achieved if social media data are combined with a market prediction model. Moro et al. (2015) and Aziz et al. (2022) collected textual data from the documents and found the important topics in the context of finance. Other machine learning techniques can also be used with the text data. Nyman et al. (2021) evaluated the risks to financial stability by using vector autoregression

model(VAR) with the estimated sentiment based on text data. Fu et al. (2020), Qian et al. (2019), and Xu et al. (2020) utilized various machine learning models such as long-short term memory(LSTM), naïve Bayes, Support Vector Machine(SVM), and clustering.

2.3.2. Text Mining Cases in Finance

Research using text mining has been actively conducted across various fields. It is especially widely used in finance because text mining can provide various financial information from textual data, such as central bank communication, news articles, or posts on social network services.

As text mining allows analysis of textual data in a quantitative way, it is able to explain and evaluate current economic status or predict uncertainty in financial markets. Bruno (2017) presented a framework for quantitative evaluation by analyzing the Bank of Italy's Financial Stability Report. Bennani (2018), Carney (2013), and Valles and Schonhardt- Bailey (2015) showed that it is possible to predict uncertainty in fiscal and economic systems and take appropriate advance measures by deriving clusters of similar or diverse topics from a central bank's minutes.

Choi et al. (2015) tried to predict corporate defaults by analyzing news articles. After selecting the 50 words that are found most frequently in articles indicating corporate defaults, they performed a decision tree analysis and found that the probability of an actual default was about 80.9% when a default keyword was included.

Song (2016) conducted a text mining analysis of social media posts related to health and welfare policies. The study identified positive signals for government policies by applying sentiment analysis to each health and welfare field. By predicting the signals required to derive future policy directions, the study showed that text mining techniques can be used for predictive analysis rather than only for interpreting a given situation. It also showed that the governments' financial execution could be evaluated by analyzing textual data about government policies. These studies confirm that text mining can provide new aspects of information to measure, evaluate or predict trends in finance.

Text mining also enables researchers to obtain inherent textual information that would be hard to analyze efficiently. Binette and Tchebotarev (2019) analyzed the Central Bank of Canada's Monetary Policy Report (MPR) to assess the readability of the text and showed that the language used in the MPR is rather too complex for the average Canadian to understand. In addition, analyzing the sentiment in the MPR showed that major events, such as the previous international financial crisis substantially influenced sentiment. Finally, word frequency and sentiment analysis are expected to help central banks draw conclusions that better reflect the actual situation. These results are difficult to obtain because significant effort by professionals to analyze qualitative data is required.

In addition, text mining can provide interesting facts in financial texts. Jang et al. (2016) analyzed the titles and forecast sections of investment strategy reports prepared by financial analysts and revealed that financial analysts' predictions are more accurate when the title contains a clear expression. Cho et al. (2018) derived open innovation research topics through topic modeling techniques and analyzed the network among these topics. They conducted topic modeling using the abstracts of related research by comparing the main topics mentioned in the research with the policies implemented by the government. They showed that it is possible to distinguish between

current policies and policies that are not being handled.

Moreover, text mining can provide more sophisticated information on central bank policy (Hansen et al., 2018; Schonhardt-Bailey, 2013) and can determine whether the mutual impacts of the various policies to be legislated are complementary or conflicting (Li et al., 2015).

Japan studied the impact of Bank of Japan's meeting minutes on the financial markets using text analysis. The study used topic modeling to derive the topics covered in the minutes, and these topics were compared with the market reaction at the time that the minutes were issued. When discussions on ways to support Japanese companies, such as negative interest rate policies, were held, stock prices declined, suggesting that the market is paying greater attention to experts' opinions on the current economic situation.

Lee et al. (2019) analyzed the minutes of the Monetary Policy Board of Korea and found that word-based indicators are helpful in explaining the Bank of Korea's monetary policy decisions. This study showed that the predictive and explanatory power of fluctuations in the base rate can be improved when the sentiment index is included as an explanatory variable in the Taylor rule, a formula used by central banks to set nominal interest rates.

Text mining is not only utilized independently but also incorporated into other methodologies, especially statistical or economic approaches. Wang et al. (2013) studied risk prediction using sentiment analysis in finance. They compare ranking analysis with regression analysis and conclude that emotional words are closely related to the risk of companies. Rekabsaz et al. (2017) predicted volatility using financial disclosures using both text features and market features through a generalized autoregressive conditional heteroskedasticity (GARCH) prediction model. They showed that their model combining text features derived from sentiment analysis outperforms the state-of-the-art models. Kumar and Ravi (Kumar and Ravi, 2016) investigated various research papers in the financial domain and showed that FOREX rate and the stock market could be predicted using text mining. In this study, it is shown that text mining is widely used to solve financial problems, including economic prediction.

From these studies, it is expected that financial studies will benefit from text mining techniques in many ways.

Ⅲ. Text Mining Techniques for Financial Information

3.1. Preparation of Financial Textual Data

3.1.1. Collecting Textual Data

The most important step prior to text mining is selecting and collecting textual data. The results obtained from text mining are not derived from arbitrary texts but rather from selected and collected texts that can provide the desired results. A common expression in computer science is "garbage in, garbage out," which means that outputs (i.e., the results) are unreliable unless appropriate inputs are used for analysis. Thus, text mining studies carefully review and select data that fit their purposes.

The targets of text mining analysis are selected from an expansive range of texts, including news articles, SNS posts, Internet community posts, research reports, government publications, and minutes. <Table 2> shows examples of textual data analyzed by different studies. Recent studies have extracted text from images and have converted audio data into

<Table 2> Financial Studies Using Textual Data

Data	Authors, Publication Year	Objective
News Article	Huang et al. (2017)	This study predicted companies' sales trends by combining the results of text analyses of financial news articles related to seven personal computer manufacturers in Taiwan and the characteristics calculated by an autoregressive integrated moving average model based on previous sales data.
Minutes	Oshima and Matsubayashi (2018)	This study analyzed the Bank of Japan's meeting minutes to examine the impact of its communications on the fiscal market.
SNS Data	Utami and Luthfi (2018)	The study investigated public opinion on Indonesia's tax system by analyzing posts on SNSs, such as Twitter and Facebook.

text for analysis, meaning that the potential range of data text mining is continuing to expand.

3.1.2. Preprocessing Textual Data

Before examining a text, it is necessary to process it to facilitate analysis. Because text takes the form of a human language, it must be converted into a form that can be processed by a computer. This process may vary depending on the purpose of the analysis. Nouns are the most important elements when trying to grasp the main subjects of a text. Verbs and adjectives must be considered to analyze sentiment in textual data; however, if the target is text generation or machine translation through machine learning, as in the case of a chatbot that understands conversations with humans, a more sophisticated form of language processing is required.

3.2. Keyword Analysis of Financial Textual Data

3.2.1. Frequency Analysis

Frequency analysis is a fundamental method for analyzing textual data based on the frequency at which words appear in a document. Word frequency is the basic form of frequency analysis, and it uses pairs of data consisting of words and their frequencies. Using this information, researchers can draw various conclusions, such as the main idea or keywords in the document. Although word frequency is a simple method, counting all of the words in a document makes it difficult to interpret the results correctly. For example, useless words, such as "the" and "a," occur very frequently. To overcome this problem, researchers who use text mining often build a collection of vocabularies of interest, called a corpus, to focus on the research topic. The results of text mining analyses may vary depending on the quality of the corpus; thus, it is necessary to construct a high-quality corpus to avoid distorting the results. Other methods are used to count meaningful words or phrases, such as the TF-IDF, n-gram, and co-occurrence methods. <Table 3> and <Table 4> describe frequency analysis techniques and cases.

3.2.2. Social Network Analysis

Social network analysis is a type of exploratory evaluation based on the relationship between network members, and it aims to understand a network's structure. This analysis allows a researcher to find key nodes that produce and spread important in-

<Table 3> Text Mining Techniques Related to Frequency Analysis

Methods	Description
n-gram	An n-gram is a group of n consecutive words. This method is widely used because analyzing a group of words provides richer information than counting the frequencies of simple words does. A combination of methodologies for evaluating word groups, such as pointwise mutual information, is used.
Co-occurrence	The n-gram method can only identify the relationships of adjacent words according to the given window size of the text. Because the n-gram method considers the order of combined words, it identifies relatively low frequencies. In contrast, co-occurrence does not consider word order but rather counts all the relationships between words that appear simultaneously within the window size of the text.
TF-IDF	Word frequency analysis is very simple and convenient for deriving the main content of a document set, but it has the disadvantage of not reflecting specificity. TF-IDF is a methodology that places higher weights on words with high specificity in a document set (Jones, 1972).

<Table 4> Financial Studies Using Frequency Analysis

Methodologies	Authors, Publication Year	Objectives
n-gram, TF-IDF	Mihalyi and Mate (2019)	This study analyzed the relationship between the frequency of words of interest and policy implementation and derived the relative frequency using TF-IDF.
Co-occurrence	Shirata et al. (2011)	The study distinguished between bankrupt and non-bankrupt companies based on the frequency of certain expressions that coincide with such words as “dividend” and “retained earnings”.

formation and understand the relationships between nodes. Social network analysis is also widely used in the financial sector. Mao et al. (2015) analyzed the relationships among bank customers using social network methodology to identify key customers in a bank’s customer network.

Social network analysis can be utilized in text mining. Social network analysis using text mining extracts information from the text and applies social network analysis methodologies by identifying co-occurrences of words. Through these co-occurrences, it is possible to obtain potential information from the relationships between words that cannot be easily identified using only word frequencies. By conducting social network analysis using co-occurrences, it is possible to deduce the relationships among words in a document and scientifically determine which words have greater influence.

In social network analysis, the concept of centrality is very important as it describes the most important node in the network. Core information about a network can be deduced by analyzing whether a specific node in the network influences neighboring nodes and holds an independent position. There are several ways to determine centrality, including connection information centrality (Freeman, 1978; Nieminen, 1974), proximity centrality (Beauchamp, 1965; Freeman, 1978; Kwahk, 2014; Sabidussi, 1966), mediation centrality (Anthonisse, 1971; Freeman, 1978), eigenvector centrality and PageRank (Zafarani et al., 2014), and beta centrality.

There are a few text mining studies in finance conducted using social network analysis. Rönnqvist and Sarlin (2015) deduce the relationships between

banks and calculated centrality by analyzing the names of banks that are mentioned together in financial news articles. According to the study, the centrality measure is not biased by common knowledge or superficial facts. Forss and Sarlin (2016) try to analyze the connectivity of companies in financial news using centrality measures and sentiment analysis. In this study, it is shown that discovering the companies that are the market movers can be achieved by centrality analysis on financial news articles. In addition, they combine sentiment analysis with information centrality ranking to analyze the mutual influence of companies, as well as companies that are not affected directly by news articles. From these studies, it is clear that centrality analysis of financial documents can deduce the close relationship among entities without any biases.

3.3. Natural Language Processing Application Using Machine Learning for Financial Textual Data

Deep learning is leading to outstanding performance in various fields, especially improvements in natural language processing. Because text mining is based on natural language processing, various deep learning technologies have naturally been applied to text mining in recent years. Otter et al. (2020) described the current status of natural language processing using deep learning. We summarize applications of natural language processing using deep learning in six areas: text embedding, document similarity analysis, document classification, document clustering, text summarization, and question answering. We explain each application's use for financial information, focusing on deep learning-based natural language processing techniques that are applicable to text mining.

3.3.1. Language Modeling and Text Embeddings

When processing text using a computer, it is necessary to change the language to make it easier for the machine (computer) to process rather than expecting the computer to understand human language. This method is called embedding. Embedding is the transformation of text into numbers or vectors to enable the use of statistical approaches, including machine learning techniques.

Although word embedding is essential to process language using a machine, some characteristics of the language are lost during transformation. For example, the bag-of-words model, which is the simplest representation of textual data, keeps a set of words and their multiplicities without retaining grammatical information, such as the word order (Harris, 1954). For this reason, researchers have devised various methods suitable for machine language processing that preserve the information in the text as much as possible.

Salton et al. (1975) proposed an information indexing technique using a vector space model for textual data in information theory, and Bengio et al. (2003) proposed a distributed representation of words based on a statistical approach. Subsequently, Google developed Word2Vec, an innovative embedding technique to convert a set of words into a vector of a given size without losing much of the information in the textual data. Recently, several word embedding methods, such as FastText, GloVe, and Swivel, have been developed. In addition, various revised embedding techniques, including text embedding (Lee, 2019) based on words as well as sentences, have been proposed, including Doc2Vec, LDA, ELMo, Transformer Network, and BERT (Mikolov et al., 2013).

3.3.2. Document Classification

Document classification and clustering are key techniques in text mining because these can provide effective and efficient approaches to distinguish documents. As classification and clustering methodologies are fundamentally based on similarities between documents, finding similarities is a primary procedure for classifying documents according to a given label or clustering them into a certain number of groups determined by the researcher.

A representative text mining technique for classifying documents is the naïve Bayes method. This method is an algorithm that is mostly used in text classification by applying a statistical technique. It classifies documents using conditional probabilities under the assumption that each word's occurrence is independent. The naïve Bayes method is known to derive appropriate results even if the independence assumption is violated, and it is widely used because it is simple and enables speedy computation (Domingos and Pazzani, 1997; Rennie et al., 2003; Zhang, 2005).

Recently, to improve the performance of document classification, deep learning-based methods have been studied extensively. In the early stage, RNN-based techniques, including LSTM and bi-LSTM were widely used, focusing on the fact that a document is a sequence of words. However, Kim (2014) showed that by using convolutional neural networks (CNN), a technique originally used for images, for sentence-level classification, a simple model and less hyperparameter tuning could achieve better performance than conventional techniques. Yang et al. (2016) proposed a hierarchical attention network that performs stepwise attention on sentences and words. The network adopted a hierarchical structure to utilize the hierarchical structure of documents on document classification. Since then, attempts have been made to classify documents using deep learning models based on language modeling. As a result, BERT is now showing state-of-the-art results (Adhikari et al., 2019).

Document classification is generally considered to be supervised learning because the model is trained using data with categories that are set in advance to predict the categories to which new data belong. However, this method cannot be used to classify documents in the absence of pre-trained data. Thus, if classification labels for each document are not provided, it is necessary to use a document clustering technique, as such techniques are based on unsupervised learning. Document clustering is highly dependent on the result of text embedding because it is conducted based on similarities among words or documents. K-means algorithm and topic modeling are generally applied to document clustering and it classifies the document into a predefined number of clusters by minimizing within-distance among embedded texts.

In finance, document classification and clustering have various uses, such as fraud detection, which categorizes abnormal implementations of financial transactions; identifying areas for governmental financial support based on social views of national policies. For example, Anand et al. (2020) indicated that it could be challenging for corporate lending to process the large amount of Loan Application Process (LAP) documents. They use deep learning techniques to classify those documents. Similarly, Kulathunga and Karunaratne (2017) suggested domain-specific clustering techniques for huge quantities of financial documents to retrieve important information. In this study, it is shown that incorporating external knowledge could improve the results of clustering for domain-specific documents.

3.3.3. Sentiment Analysis

Sentiment analysis uses a document classification approach to identify sentiment within a document. This method can easily deduce the tone, sensibility, and attitude of a document, which are difficult to identify using traditional methodologies. Many studies using sentiment analysis are being actively conducted by combining emotional information with economic methodologies, and in many cases, the prediction accuracy of the existing methodology is significantly improved as a result.

Sentiment analysis requires a predefined sentiment lexicon that indicates the emotional polarity of each word, especially in the proper context. Therefore, financial studies using sentiment analysis are mainly dependent on the sentiment lexicon. Li and Shah (2017) suggested a novel approach to building a sentiment lexicon from a social network to study stock markets. They use machine learning techniques to incorporate sentiment information into word embedding that contains syntactical contexts. Keith and Stent (2019) used financial sentiment lexicons to analyze the sentiment of earnings calls. Importantly, they incorporate two bundles of lexicons, one for general purpose and the other for financial study, to analyze both formal and informal statements.

3.3.4. Topic Modeling

There are various ways to cluster documents, but the most common methodology is topic modeling, which can deduce sets of words that are closely related to each other in terms of a topic or an issue. In other words, each set of words derived from topic modeling represents a topic mentioned in a document.

The most popular algorithm for topic modeling is Latent Dirichlet Allocation (LDA) which classifies words based on the predefined number of topics (Blei et al., 2003). As LDA uses the unsupervised method to estimate the set of words, supervised methods such as supervised LDA and labeled LDA, are also devised (Mcauliffe and Blei, 2007; Ramage et al., 2009).

Topic modeling is widely used because it can quickly derive the main topics from a set of documents. For example, when analyzing a large number of social media posts about government policy, it is difficult to identify the main topics being mentioned. However, topic modeling allows the researcher to cluster words together according to related topics in an innovative and fast way. Topic modeling also helps to identify issues that have not drawn much attention because its results are derived not from the researcher's bounded knowledge but from statistical inferences, which generate impartial results.

The results of topic modeling can also be used to investigate trends that change over time. This approach is referred to as topic trend analysis (Cho et al., 2017). Topic trend analysis allows a researcher to observe dynamic changes in the topics mentioned in documents over time and identify the lifetimes of specific issues. For example, Si et al. (2013) studied stock prediction using topic-based Twitter sentiment over time. They first estimate the number of topics using the extended LDA model, called the continuous Dirichlet processes mixture (DPM) model, and incorporate the score of emotional polarity into the topics.

3.3.5. Text Summarization

Text summarization refers to finding important parts of a given document and summarizing them into smaller texts, which can be used when a researcher does not have enough time to read long documents. In the case of financial information, text summaries

can save time and provide more accurate results than reading a document to understand its content can. Because of its usefulness, there have been some efforts to summarize the financial text automatically for a long time. de Oliveira et al. (2002) present a system to summarize financial news based on lexical cohesion. Yang and Wang (2003) suggest a similar automatic summarization system for financial news articles to overcome the physical limitation of mobile devices at that time. Thus, text summarization can be achieved by text mining techniques such as co-occurrence analysis or TF-IDF. Also, even though past studies sought to summarize financial news articles for general purpose, it is expected that summarization for financial documents or news articles could be beneficial to resolving recent problems such as the need for taking immediate action to stock market variability.

Rush et al. (2015) encoded the contextual information of sentences in a document using a convolutional attention-based encoder and created a summary using a generative beam search decoder. This study is significant as it showed that performance using deep learning models is comparable to that using state-of-the-art models. As in the case of document classification, the model using BERT currently performs the best Zhang et al. (2019). Therefore, text summarization is promising in text mining for financial studies and practical applications.

3.3.6. Question Answering

Question answering automatically generates an answer to a given question by learning about documents composed of existing question-answer pairs. The basic form and the fundamental approach are similar to text summarization or machine translation, except that the contents are classified by questions and answers as inputs and outputs. The use of chatbots to automatically answer questions is increasing, and this service may be important in the field of financial information in the future. It is difficult to say that question answering based on deep learning performs sufficiently, but its performance can be greatly enhanced if natural language processing technology improves and the data necessary for question answering learning accumulate. Thus, this application field is much wider than those of other deep-learning-based text mining tools. Wang et al. (2017) generated answers by matching sentences, including a question and an answer, using a gated attention-based recurrent network. Dong et al. (2015) analyzed questions using conventional neural networks and ranked possible answers. As with other methodologies, the state-of-the-art model uses BERT. Yang et al. (2019) obtained the best results so far in a study on providing appropriate answers to a given problem using Wikipedia articles.

Ⅳ. Empirical Case Study of Text Mining in Financial Information

Government policy includes preparing and executing a budget. In other words, an analysis of government policy can study the selection of national fiscal execution targets, budget formulation, and the evaluation of budget execution results. Thus, this study primarily analyzes news articles on the South Korean national policy to apply text mining to the financial industry.

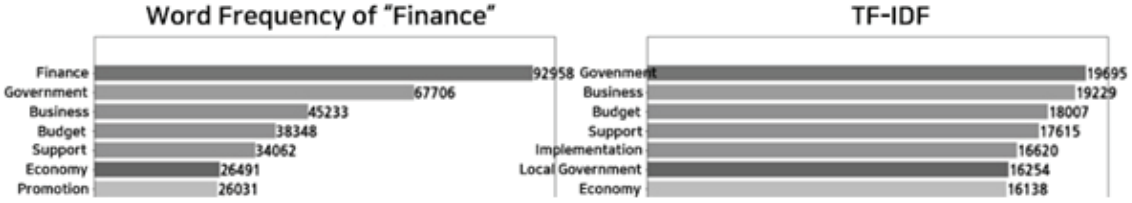
4.1. Text Mining Applications to "Government Finance" from News Articles

We collect and analyze news articles from the first quarter of 2010 to the fourth quarter of 2019 that

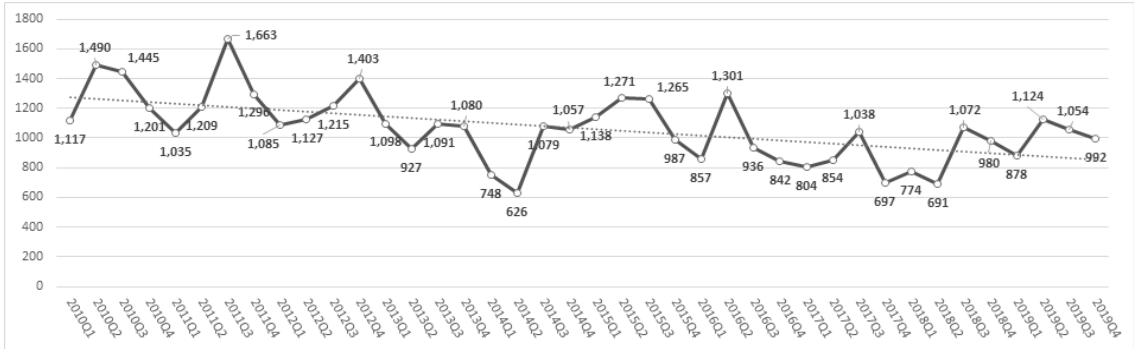
include the keyword "finance" in their titles. News articles with “finance” in the title describe a wide range of financial activities required to execute government policies from a national perspective as well as corporate, local government, and international finance. Thus, an analysis of news articles with “finance” in the title is expected to show the overall status of finance in South Korea. This study obtains about 50,000 news articles and analyzes 42,547 of them, excluding those that differ from the context of this analysis.

<Figure 2> shows the word frequency and TF-IDF results for the entire document. The most frequent words are “government,” “local government,” “Ministry of Economy and Finance,” “National Assembly,” and “president,” which are related to government agencies and “project,” “support,” “execution,” “promotion,” and “plan,” which are related to government activities. These results arise because, in the news, “finance” mainly refers to economic activities for implementing national policy. In addition, words such as “economy,” “evaluation,” “situation,” and “market” represent the economic statuses of households, companies, and governments; thus, it can be seen that “finance” refers to the national economy in the obtained data.

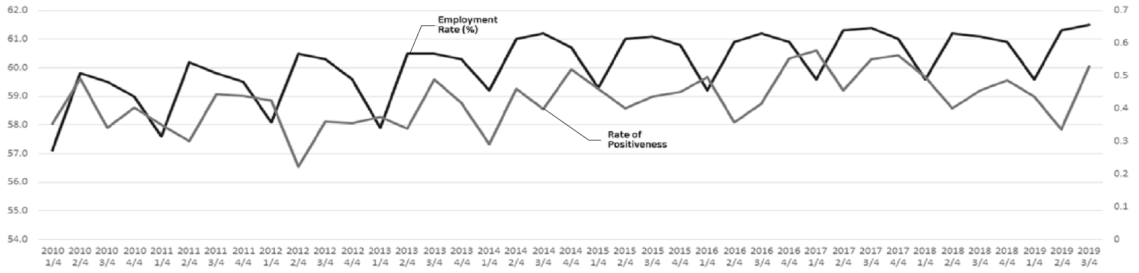
Although <Figure 3> shows that the number of news articles seems to be gradually decreasing, it is difficult to determine whether interest in finance has declined because the impact of “finance” on society is not small and public opinion’s main concern is the policy by which “finance” is actually implemented. However, this result indicates that a better understanding of national financial conditions among the general public is needed to reduce discord in the national decision-making process, as policies are established and executed according to the sizes and states of financial, economic, and social circumstances.



<Figure 2> Word Frequency and TF-IDF Results for News Articles with “Finance” in the Title



<Figure 3> The Number of News Articles with “Finance” in the Title



<Figure 4> Comparison of Sentiment Analysis of Employment Rate and Job/employment Policy

The main ideas covered in the news articles are closely related to key national policies, and government policy generally focuses on major social issues, meaning that the main topics comprise social issues and content related to the government. We applied the latent Dirichlet allocation (LDA) model for topic modeling on the news articles. For the result, we derived 60 initial topics and chose 25 main topics from those based on the probability of each topic (see <Table A> in <Appendix>). In the model, we set the parameter of the Dirichlet prior on the per-document topic distributions, alpha, as 0.1 and the parameter of the Dirichlet prior on the per-topic word distribution, beta, as 0.01. <Table 5> shows the main topics derived from the news articles by using topic modeling.

The main topics related to finance in <Table 5> are policies that generally relate to fiscal enforcement. For example, "employment/welfare," "education/welfare," "health/welfare," "pension/welfare," "job/employment," and "education/policy," which require large budgets, account for about 24% of the topic modeling results. This result is not significantly different from the 34.2% share of health, welfare, and employment in Korea's 2019 budget.

Each topic was derived from all articles released from 2010 to 2019. From the changes in public opinion on each topic over time, we can evaluate the appropriateness of the policy direction or determine the public opinion in response to a specific policy. <Figure 4> compares the emotional trend (positiveness) for "job/employment" derived from the given data with

<Table 5> Results of Topic Modeling

Topics	Keywords
Economy/Finance	Economy, exports, growth rate, economy, interest rate, growth, Korea, Bank of Korea, forecast, governor, fiscal policy, slowdown
Education/Policy	Private high school, student, education, school, selection, high school, general high school, selection, SAT, application
North Korea	North Korea, North and South, Korean Peninsula, unification, security, peace, diplomacy, provocation, missile
Pension/Welfare	Pension, national pension, public employee pension, basic pension, estimate, public official, reform, reform proposal
Job/Employment	Jobs, employment, income, minimum wage, support, small and medium-size enterprises (SMEs), wages, creation

trends in South Korea's employment rate. In this example, we use the sentiment lexicon for general purposes, as proposed by Park et al. (2018). Using this lexicon, we could derive a sentiment polarity score of each news article and calculate the average score of positiveness for articles issued in the same period (see <Table B> in <Appendix>).

The positive sentiment of public opinion related to jobs/employment in <Figure 4> shows a similar pattern to the trend in the employment rate. In other words, the sentiment of public opinion on jobs/employment shows a similar trend to an actual economic indicator.

In this example, news articles related to "finance" are analyzed using text mining techniques to derive keywords, and from that result, the change in the number of documents over time is analyzed. In addition, major topics are derived using topic modeling, and changes in the positive sentiment around "job/employment" and the employment rate over time are investigated simultaneously.

4.2. Text Mining on "Free Education" from News Articles

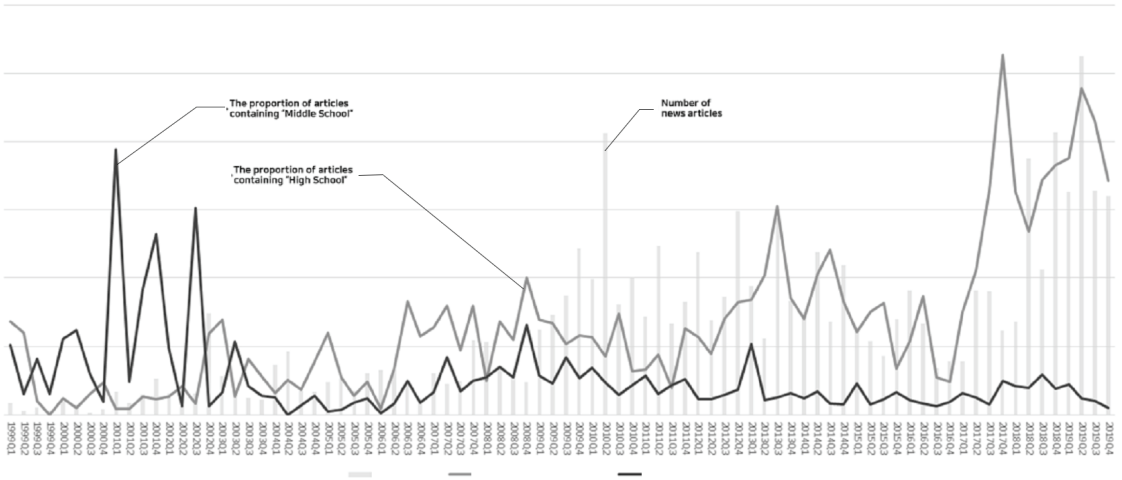
According to "Public Finance of Korea 2019" published by the National Assembly Budget Office, health, welfare, and employment account for 34.2% of the 2019 budget (approximately USD 135 billion out of USD 400 billion). Thus, policies regarding health, welfare, and employment are accorded utmost importance in South Korea (NABO, 2019). If the scope of welfare is extended to education, environment, and safety, its share will amount to 50% of the total budget.

As welfare requires a large budget, policies must be designed to resolve social issues. Text mining analysis can be used to investigate whether the proposal and implementation of welfare policies are appropriate based on the social issues in South Korea.

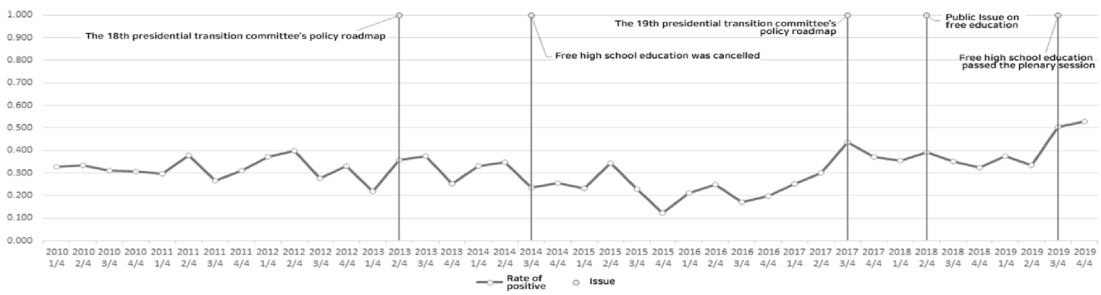
<Figure 5> shows that a large proportion of news articles about free education focus on free high school education. According to the results of frequency analysis, free high school education has been a representative issue in Korea in recent years and has been a major issue for a long time.

Because free middle school education is provided in Korea, the number of articles that mention "high school" and "free high school" education has been increasing, and the issue became a major issue in 2013. In fact, the Ministry of Education in South Korea announced a plan to implement free high school education sequentially starting in 2014 on the basis of the state affairs of the 18th Presidential Transition Committee. On May 31, 2013, the Ministry of Economy and Finance pledged to provide free high school education in the financial plan for future policies, which describes the financial resources needed to implement free high school education (MOE, 2018). However, because free high school education was canceled for some reason, mentions of free high school education exhibited a decreasing trend. Nevertheless, <Figure 5> suggests that the ongoing debate over free high school education has continued over the years. Free high school education once again received an explosive public response after the "Five-year Roadmap" in the 19th presidential election was unveiled in 2017. In 2019, the free high school education bill was finally completed with the passage of a plenary session of the National Assembly.

The results of the sentiment analysis of public opinion regarding free high school education are shown in <Figure 6>. The figure shows that public opinion changes over time according to government activities, and positive sentiment regarding free high school education increased when it finally passed the plenary



<Figure 5> Number and Proportion of News Articles Containing "High School" and "Middle School"



<Figure 6> Changes in Public Sentiment Regarding Free High School Education
(Higher Scores Imply More Positive Sentiment)

session of the National Assembly.

Ⅴ. Topics for Future Financial Research Using Text Miming

Text mining is a technique for dealing with unstructured text data. It has the advantage of accurately describing actual phenomena by avoiding the researcher's bias that appears in the analysis process. In addition, it enables systematic analysis based on verified statistical methodologies and provides the ability to obtain useful information that was previously difficult to find. In the field of financial information, text mining is used as a tool for analyzing and evaluating current and past phenomena as well as for predicting future phenomena.

Keywords related to a specific topic can be checked over time, making it easy to identify differences in texts before and after a phenomenon to be observed. For example, by analyzing news articles, social media posts, or texts from online communities to derive the most frequently cited content related to the establishment of a local transport policy, a researcher can easily observe the actual demand and requirements related to the policy. As in the example analysis of

free education discussed above, trends in keywords from the past to the present can be analyzed over time, and long-term and short-term changes in these trends can be identified before and after actual policy changes. The results of analyses across data from the past and present can serve as the basis for designing and predicting the future. The results of public opinion surveys on policies not only help in identifying policy demands and establishing correct policy directions but also enable fair budget allocations and appropriate finances to be secured in the future. Thus, it is necessary to judge accurately past and present situations and prepare for the future, and this goal can be achieved through text mining techniques. The research topics that must be addressed in the field of financial information with text mining are listed in <Table 6>.

First, building a domain-specific corpus is very important to discover various aspects of the text in a specific area of research because the result of text mining is highly dependent on textual data. A sentiment lexicon for sentiment analysis in finance is also necessary. Without a corpus or lexicon prepared for the financial field, it is difficult to obtain robust results of financial research using text mining, and furthermore, it might draw an unexpected conclusion that is different from the real world. According to the Loughran and McDonald (2011), the sentiment lexicon for the general purpose could mislead the text mining result in financial texts. However, there are a few studies in building a corpus, and a lexicon for the financial sector is required for further studies using text mining techniques. It can be achieved by constructing domain-specific dictionaries by con-

<Table 6> Proposed Research Topics Using Text Mining with Financial Information

No.	Subject	Description
1	Build a Corpus of Financial Information	- A corpus for text mining research in the field of financial information should be built. - Once a financial information corpus is established and disclosed, various text mining studies related to financial information can be conducted.
2	Detect Financial Execution Anomalies	- Incorrect financial execution is detected by identifying instances of budget wasting by government agencies, including local governments, and cases of policy execution in a different form from the plan, using news articles or SNSs. - Transparency in financial execution can be enhanced by increasing the number of public opinions on financial information.
3	Evaluate Policies (businesses)	- Policy evaluation combines text analysis results and department-specific data. A financial business performance management system can be established by integrating department-specific data, text data, and analysis results. - Policy decisions and future development directions can be supported by providing project evaluation information combining text analysis results and departmental data.
4	Discover Policy Blind Spots	- Public opinions from various media can be analyzed to identify core policies that do not effectively address social issues. - Financial and policy support measures can be prepared for socially underprivileged groups with difficulty forming public opinions.
5	Provide Budgeting Support	- Combining quantitative and qualitative data can help to support local governments or government budgeting. - Effective budget deliberation is possible by minimizing unnecessary disputes and focusing on detailed budget planning discussions.

sensus by professionals or related research or making corpus or lexicon using statistical approaches or machine learning techniques.

Second, many organizations are trying to detect anomalies in finance, which might initiate a tremendous loss of money as well as credibility. In the case of public finance, it is hard to find financial execution anomalies because of the scale of money and the complexity of the system. For these reasons, those anomalies are usually discovered by news articles or are broadcast in both a direct and an indirect way. Text mining would simplify the process of finding hidden anomalies quantitatively and would also help to provide useful information to decision-makers, especially in government budgeting. Based on quantitative results, distinguishing documents that include abnormal information can be achieved using text mining techniques such as document classification and clustering.

Third, the use of text mining will help to evaluate financial execution or policies related to finance. Although public opinion should be collected very carefully to be analyzed, biased researchers could influence a survey or a questionnaire to collect the opinions and it also requires tremendous amounts of time and money. However, sentiment analysis can be used to comprehend a diversity of public or professional opinions with minimizing stereotypes and broadening the view of the issue. Indeed, it can save a lot of time and money from collecting to analyzing data as well.

Fourth, discovering policy blind spots becomes increasingly important for many reasons. This is important because recognizing unknown facts can lead decision-makers in the right direction. Text mining allows the discovery of inherent topics in textual data, which are hard to deduce without background knowledge. For example, companies can modify their policies after discovering key operational knowledge, while Government organizations can achieve social justice after discovering issues regarding unprivileged groups that it did not recognize. As it is based on a statistical approach, unexpected topics can emerge, which could be very helpful to decision-makers. Topic modeling, a novel text mining technique, can identify topics in a document automatically.

Lastly, providing budgeting support is a promising topic because correct budgeting is very hard to achieve and needs both quantitative and qualitative information to make better decisions. As text mining techniques use quantitative approaches, the result would be reliable and useful. Next, text mining can help sift through and isolate several topics or issues from tons of documents. Classification or clustering the important textual materials would be very helpful for decision-makers. Moreover, text mining can help government decision-makers to recognize the main issue of underprivileged people by providing various information from, not only statistical data, but also social media, news articles or investigation reports.

5.1. Building a Corpus of Financial Information

In order to conduct text mining analysis using financial information, it is necessary to build a corpus suitable for financial information. In the field of natural language processing, a corpus is a major resource that determines the quality of text analysis. Thus, countries with their own languages are trying to build their own corpora. The U.S. has built about 300 billion English corpora, whereas China has about 80 billion words, and Japan has approximately 15 billion words. However, to analyze various finance-related texts in English, Chinese, and Japanese, it is necessary to establish a corpus suitable for the field of financial

information. When such a corpus is established, unnecessary pre-processing procedures can be reduced, helping to prevent the distortion of results and thereby creating the necessary conditions for various studies.

5.2. Detecting Financial Execution Anomalies

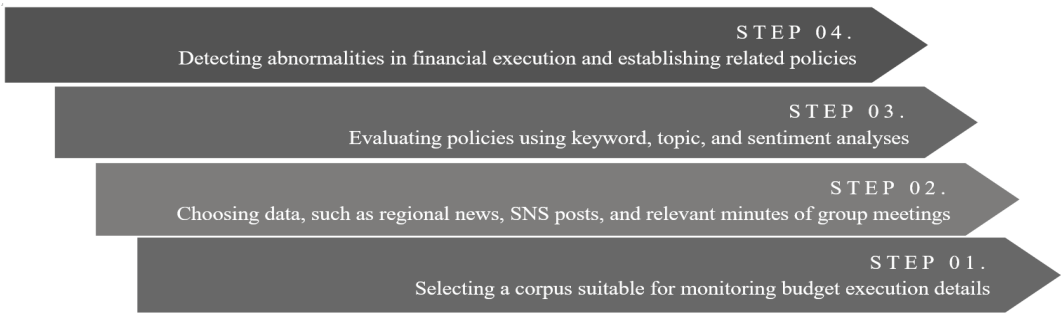
Fraud detection is a technology that attracts attention in the field of big data analytics, and it is used in various fields, such as industrial sites, network management, and advertisement services. Specifically, it is widely used in the financial sector to facilitate the detection of fraudulent card use. For example, the U.S. Treasury Department continues to refer to text mining or text analysis in its Annual Privacy Report, meaning that text mining is necessary for analysis. Moreover, the U.S. House of Representatives also recently passed a bill to continue research using such technologies as text mining and blockchain in the Financial Crimes Enforcement Network (USDT, 2015).

In budgeting, anomaly detection also can be used to identify cases in which a budget is not properly executed owing to wasted money or incorrect demand forecasting, thereby creating the necessary structures for responding at the appropriate time. Once a system that monitors and evaluates whether a budget is being properly executed using regional news, SNS posts, or the minutes of related organizations or local governments is established, unnecessary budget allocation waste can be prevented, and transparency in budget execution can be secured. <Figure 7> shows the steps involved in the anomaly detection system for financial execution.

5.3. Evaluation of Policy Programs

One of the biggest advantages of text analysis is that it can analyze public opinion efficiently and effectively while minimizing the distortion caused by the researcher's bias. For instance, the results of implemented public welfare policies can be evaluated through public opinion. Using topic modeling, it is possible to identify the main social issues to compare with current policies to determine whether those issues are already properly detected and treated by the government. Furthermore, by performing sentiment analysis on each topic, policies can be evaluated by estimating the public's positive and negative sentiments.

Specifically, if a performance management system combines qualitative data and text information and uses these data to measure and evaluate performance, it not only helps in organizing and executing effective budgets but also enables correct decision-making as an indicator of policy decisions and directions for



<Figure 7> Example Anomaly Detection System for Financial Execution

future development. As a result, a performance management system that uses textual information will be able to meet internal and external demands more accurately.

5.4. Discovering Policy Blind Spots

Text mining can derive information on major issues and simultaneously detect non-mainstream public opinions. For example, when deriving information on a topic covered by public opinion through topic modeling, not only do issues that are mentioned at a high frequency stand out, but important topics that are not found in the flow of public opinion can also be discovered. Specifically, text mining is believed to be of great help in addressing "policy blind spots" for the socially weak.

South Korea's Social Security Information Service (SSIS) operates an information system with the aim of identifying welfare blind spots and efficient financial management using data from each ministry in various ways. The SSIS plans to develop welfare services from "letting people apply" to "finding people in advance" by automatically identifying the socially underprivileged by 2040. In 2018, more than 130,000 people in welfare blind spots and 50,000 children suspected to be crisis victims were identified and successfully supported by SSIS.

As such, it is necessary to support budgeting and policy establishment for socially disadvantaged people who fall into policy blind spots. Furthermore, this activity should be carried out across other fields, such as education, defense, local government, and employment. Text mining techniques, such as topic modeling, sentiment analysis, and social network analysis, can allow the government to achieve the goal of identifying the socially underprivileged by discovering policy blind spots in an innovative way.

5.5. Support for Budget Planning

Monitoring and evaluating budget and policy execution through text mining can enable efficient execution of the budget, proper implementation of policy, and necessary policy establishment, reducing inefficient work that requires substantial time and money. For example, discussions and disputes without compromises on policies and interminable confrontations in budget planning and execution often lead to inefficiencies in policy programs. Time and money are therefore wasted on inefficient debates even though in-depth analyses of more detailed content are required to discuss numerous policy projects efficiently; as a result, appropriate budget execution cannot be implemented in a timely way.

Recently, services that support automated budgeting have emerged to improve transparency, effectiveness, and efficiency in government budgeting (OPM, 2021). An automated budget system can achieve a high level of transparency in managing the budgeting process and can foster close interactions between various ministries to enable appropriate budget allocation and execution (Accenture, 2014). Following this trend, text mining techniques can be used to identify public opinion, discover policy blind spots, and calculate the budget. At the same time, if quantitative data analysis and machine learning techniques are performed simultaneously, it is believed that basic reference materials necessary for budget planning can be derived.

The use of text mining in budget planning can benefit from clearing blind spots by identifying policy demands, discovering recent trends and issues in public opinion, presenting basic policy suggestions, and enabling in-depth discussions on detailed policies. In addition, because this process proposes a framework for budget proposals based on analyses of past

budget proposals, current issues, and the foreseeable future, it is expected to enable more transparent budget planning and execution and policy establishment and execution.

Ⅵ. Conclusion and Contribution

6.1. Conclusion

Text mining is a technique for analyzing textual data, which accounts for about 80% of total data. It has the advantage of providing richer information relative to general data analysis. In addition, potential key features of text mining include data unbiasedness, efficiency, effectiveness, and the possibility of deriving new information.

Although text mining is widely used across most fields and provides various advantages, text mining in the field of financial information is still in an early stage compared to other fields. However, it is encouraging that several studies in related fields are being conducted overseas, and institutions and companies, such as countries' national banks, government organizations, and corporations, are trying to apply text mining techniques to improve their analyses and obtain informative results.

Building a corpus of financial information is necessary to encourage text mining studies in finance. Several goals can be achieved using text mining techniques, such as detecting financial execution abnormalities, evaluating policy projects, and discovering policy blind spots. In addition, it is possible to conduct research to support organizations in budget preparation using machine learning techniques and to improve economic models in the financial sector by combining various additional pieces of information derived from text mining.

6.2. Discussions and Contributions

This study provides a literature review, as well as practical examples that help to understand the benefit of text analysis and how text mining techniques can be applied in the financial sector. We describe text mining techniques and related studies to help understand how it works and what has been conducted. It is important to review text mining studies in finance because this would encourage researchers or practitioners to use text mining techniques. Aziz et al. (2022) also pointed out that only a few applications of machine learning for financial problem have been published in finance journals. Likewise, there are still many financial problems to be enhanced using text mining applications. We believe that our review of text mining in finance will spur the interest of researchers to conduct more developed studies.

This study has both academic and practical implications. From an academic perspective, it presents a comprehensive review of the literature on text mining to broaden understanding of the major issues related to textual analysis in the financial sector. In addition, textual data, such as online posts, comments, news, and even words in video and audio sources, can provide a wealth of information that can lead to new insights in finance and fiscal policy. Moreover, several text mining technologies that are mainly used in this sector are discussed to support the selection of methods affiliated with the study objectives.

From a practical perspective, our work presents several studies and reports to provide deeper insight into the practical applications of text mining technologies. As this review widens the possibility and potentiality of text mining in the field of finance, organizations can be encouraged to use this technique in a variety of contexts. In addition, practical examples

of text mining in Section 4 are a guide to using technology in the financial sector. Based on a review of the relevant literature and text mining technologies, we also suggest future research directions and tasks, as well as its necessity and feasibility to apply text mining to analyze data in the fiscal policy in Section 5. It would be a sound blueprint for the practitioners who want to broaden the boundary of their application as it could encourage the application of text mining techniques in finance more actively as a practical method.

Therefore, we believe this study could be beneficial to researchers and practitioners in finance. However, our study has the following limitations: First, the detailed impact or influence of text mining on the financial sector could be analyzed and measured in further studies. Second, in addition to the descriptive approach, it would be better to include predictive and prescriptive approaches to the example of text mining applications in future work. Third, the result of practical examples could be qualified, and these should be verified carefully to apply to the real world. Lastly, this study only focuses on the usefulness of text mining in financial sector and provides the practical application on fiscal policy in South Korea. In order to encourage financial research using text mining, it is necessary to first develop a corpus containing useful financial texts, and various examples of text mining.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02089039).

<References>

[1] Accenture, C. (2014). More Bang for the Budget : Automating Budget Processes for Government Efficiency, Retrieved from https://accntu.re/3u04U23

[2] Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

[3] Anand, G. S., Kuriakose, J., Sharma, S., and Guha, D. (2020, 4-7 Nov. 2020). Deep learning for information extraction in finance documents: Corporate loan operations. *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*.

[4] Anthonisse, J. M. (1971). The rush in a directed graph. *Stichting Mathematisch Centrum. Mathematische Besliskunde* (BN 9/71).

[5] Aziz, S., Dowling, M., Hammami, H., and Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *European Financial Management*, *28*(3), 744-770.

[6] Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, *10*(2), 161-163.

[7] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137-1155.

[8] Bennani, H. (2018). The art of central banks' forward guidance at the zero lower bound [La pratique des indications prospectives des banques centrales dans le contexte de la borne du zéro sur les taux d'intérêt]. *Revue économique*, *69*(1), 111-137.

[9] Binette, A., and Tchebotarev, D. (2019). *Canada's Monetary Policy Report: If Text Could Speak, What Would It Say?,* Retrieved from https://bit.ly/2QvMxUT

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

[11] Bruno, G. (2017). Central bank communications: Information extraction and semantic analysis. *Big Data, 44*. Bank for International Settlements.

[12] Carney, M. (2013). Crossing the threshold to recovery. *Bank of England Speech*, *28*.

[13] Chen, W., Lai, K., and Cai, Y. (2018). Topic generation for Chinese stocks: A cognitively motivated topic modeling method using social media data. *Quantitative Finance and Economics*, *2*(2), 279-293.

[14] Cho, K. W., Bae, S. K., and Woo, Y. W. (2017). Analysis on topic trends and topic modeling of KSHSM journal papers using text mining. *The Korean Journal of Health Service Management*, *11*(4), 213-224.

[15] Cho, S. B., Shin, S. A., and Kang, D. S. (2018). A study on the research trends on open innovation using topic modeling. *Informatization Policy*, *25*(3), 52-74. https://doi.org/10.22693/NIAIP.2018.25.3.052

[16] Choi, J. W., Han, H. S., Lee, M., and An, J. M. (2015). The prediction of corporate bankruptcy using text-mining methodology. *Korea Productivity Association*, *29*(1), 201-228.

[17] Connelly, R., Playford, C. J., Gayle, V., and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1-12.

[18] Davis, A. K., Piger, J. M., and Sedor, L. M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research*, *29*(3), 845-868.

[19] de Oliveira, P. C. F., Ahmad, K., and Gillam, L. (2002). A financial news summarization system based on lexical cohesion. *Proceedings of the International Conference on Terminology and Knowledge Engineering*. Nancy, France.

[20] Domingos, P., and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, *29*(2-3), 103-130.

[21] Dong, L., Wei, F., Zhou, M., and Xu, K. (2015). Question answering over freebase with multi-column convolutional neural networks. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, *1*, 260-269.

[22] FiveThirtyEight. (2016). Who will win the presidency?, Retrieved from https://projects.fivethirtyeight.com/2016-election-forecast/

[23] Forss, T., and Sarlin, P. (2016). From news to company networks: Co-occurrence, sentiment, and information centrality. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*.

[24] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215-239.

[25] Fu, X., Ouyang, T., Chen, J., and Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing & Management*, *57*(4), 102236. https://doi.org/10.1016/j.ipm.2020.102236

[26] Guo, H., Wang, L., Chen, F., and Liang, D. (2014). Scientific big data and digital earth. *Chinese Science Bulletin*, *59*(35), 5066-5073.

[27] Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801-870.

[28] Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

[29] Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*.

[30] Huang, H. C., Hwang, S. Y., Chang, S., and Kang, Y. (2017). Forecasting company revenue trend using financial news. *Pacific Asia Conference on Information Systems (PACIS)*.

[31] Jang, J. K., Lee, K. H., and Lee, Z. (2016). How the title of investment strategy report affects stock price forecast: Using text mining method. *Korea Bigdata Society*, *1*(2), 21-34.

[32] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21. https://doi.org/10.1108/eb026526

[33] Junqué de Fortuny, E., De Smedt, T., Martens, D.,

and Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, *50*(2), 426-441. https://doi.org/10.1016/j.ipm.2013.12.002

[34] Keith, K. A., and Stent, A. (2019). Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 493 – 503.

[35] Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Doha, Qatar.

[36] Kulathunga, C., and Karunaratne, D. D. (2017). An ontology-based and domain specific clustering methodology for financial documents. *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer).*

[37] Kumar, B. S., and Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128-147. https://doi.org/10.1016/j.knosys.2016.10.003

[38] Kwahk, K. Y. (2014). *Social Network Analysis*. CHUNGRAM.

[39] Lee, Y., Kim, S., and Park, K. (2019). Deciphering monetary policy committee minutes with text mining approach: A case of Korea. *Korean Economic Review, 35*(2), 471-511. https://doi.org/10.22841/kerdoi.2019.35.2.008

[40] Li, Q., and Shah, S. (2017). Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017).*

[41] Li, W., Azar, P., Larochelle, D., Hill, P., and Lo, A. W. (2015). Law is code: a software engineering approach to analyzing the united states code. *J. Bus. & Tech. L.*, *10*, 297.

[42] Loughran, T., and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10 Ks. *The Journal of Finance*, *66*(1), 35-65.

[43] Mao, H., Jin, X., and Zhu, L. (2015). Methods of measuring influence of bank customer using social network model. *American Journal of Industrial and Business Management*, *5*(4), 155.

[44] Mcauliffe, J., and Blei, D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*, *20*, 121-128.

[45] Mihalyi, D., and Mate, A. (2019). Text-Mining IMF Country Reports – An Original Dataset. *MPRA Paper 100656.* University Library of Munich, Germany.

[46] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 2*, 3111 – 3119. https://doi.org/10.5555/2999792.2999959

[47] Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

[48] MOE. (2018). *A study on the ways to relize free high school education*.

[49] Moro, S., Cortez, P., and Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314-1324.

[50] NABO, K. (2019). Public Finance of Korea 2019. *National Assembly Budget Office*.

[51] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653-7670.

[52] Nguyen, T. H., and Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 1354 – 1364.

[53] Nieminen, J. (1974). On the centrality in a graph. *Scandinavian Journal of Psychology*, *15*(1), 332-336.

[54] Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, *127*, 104119.

[55] OPM, C. (2021). Automated budget system, Retrieved from https://bit.ly/3wiLR4r

[56] Oracle, C. (2022). Exadata cloud increases financial services insight and agility, Retrieved from https://www.oracle.com/database/what-is-data-management/financial-services/

[57] Oshima, Y., and Matsubayashi, Y. (2018). Monetary policy communication of the bank of Japan: Computational text analysis. Discussion Papers 1816, Graduate School of Economics, Kobe University.

[58] Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21.

[59] Pang, B., and Lee, L. (2009). Opinion mining and sentiment analysis. *Comput. Linguist*, *35*(2), 311-312.

[60] Park, S. M., Na, C. W., Choi, M. S., Lee, D. H., and On, B. W. (2018). KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon. *Journal of Intelligence and Information Systems*, *24*(4), 219-240.

[61] Pejic-Bach, M., Pivar, J., and Krstić, Ž. (2019). Big data for prediction: Patent analysis – Patenting big data for prediction analysis. In *Big Data Governance and Perspectives in Knowledge Management* (pp. 218-240). IGI Global.

[62] Pejić Bach, M., Krstić, Ž., Seljan, S., and Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*(5), 1277.

[63] Qian, Y., Deng, X., Ye, Q., Ma, B., and Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing & Management*, *56*(6), 102086. https://doi.org/10.1016/j.ipm.2019.102086

[64] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 conference on empirical methods in natural language processing,* 248 – 256.

[65] Rekabsaz, N., Lupu, M., Baklanov, A., Hanbury, A., Dür, A., and Anderson, L. (2017). Volatility prediction using financial disclosures sentiments with word embedding-based IR models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 1712 – 1721.

[66] Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th international conference on machine learning (ICML-03),* 616 – 623.

[67] Rönnqvist, S., and Sarlin, P. (2015). Bank networks from text: interrelations, centrality and determinants. *Quantitative Finance*, *15*(10), 1619-1635.

[68] Rugters, U. (2015). Big data in accounting : An overview. Rutgers Business School, Retrieved from https://bit.ly/3hDLxtd

[69] Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 379 – 389.

[70] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581-603.

[71] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

[72] Schonhardt-Bailey, C. (2013). Deliberating American policy: A textual analysis. MIT Press, Cambridge, MA.

[73] Schumaker, R. P., Zhang, Y., Huang, C. N., and Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458-464.

[74] Shirata, C. Y., Takeuchi, H., Ogino, S., and Watanabe, H. (2011). Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of Emerging Technologies in Accounting*, *8*(1), 31-44.

[75] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and

Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 24-29.

[76] Silver, N. (2016). Donald Trump Has A 20 Percent Chance Of Becoming President, Retrieved from https://fivethirtyeight.com/features/donald-trump-has-a-20-percent-chance-of-becoming-president/

[77] Song, M. (2017). *Text Mining*. CHUNGRAM.

[78] Song, T. (2016). Using social big data predictive future signal: With special reference to the major policy issues of health and welfare. *Health and Welfare Policy Forum*, *2016*(8), 17-30.

[79] Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, *118*, 26-40.

[80] Stephens-Davidowitz, S. (2017). *Everybody lies*. Harper Collins.

[81] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139-1168.

[82] Turner, D., Schroeck, M., and Shockley, R. (2013). *Analytics: The real-world use of big data in financial services*. IBM Global Business Services.

[83] USDT. (2015). *Annual Privacy and Data Mining Report*. U.S. Department of the Treasury, Retrieved from https://home.treasury.gov/footer/privacy-act/privacy-reports

[84] Utami, E., and Luthfi, E. T. (2018). Text mining based on tax comments as big data analysis using SVM and feature selection. *2018 International Conference on Information and Communications Technology (ICOIACT)*, 537-542.

[85] Valles, D., and Schonhardt-Bailey, C. (2015). *Forward Guidance as Central Bank Discourse: MPC Minutes and Speeches under King and Carney* Political Leadership and Economic Crisis Symposium, Yale University.

[86] Wang, C. J., Tsai, M. F., Liu, T., and Chang, C. T. (2013). Financial sentiment analysis for risk prediction. *Proceedings of the Sixth International Joint Conference on Natural Language Processing,* 802-808.

[87] Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 189-198.

[88] Xu, F., Pan, Z., and Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management*, *57*(5), 102221. https://doi.org/10.1016/j.ipm.2020.102221

[89] Yang, C. C., and Wang, F. L. (2003). Automatic summarization for financial news delivery on mobile devices, Retrieved from http://www2003.org/cdrom/papers/poster/p178/p178-yang.html

[90] Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). End-to-end open-domain question answering with BERTserini. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Demonstrations), 72-77.

[91] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1480 – 1489.

[92] Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

[93] Zhang, H. (2005). Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, *19*(02), 183-198.

[94] Zhang, H., Cai, J., Xu, J., and Wang, J. (2019). Pretraining-Based Natural Language Generation for Text Summarization. *Proceedings of the 23rd Conference on Computational Natural Language Learning* (CoNLL), 789-797.

<Appendix>

<Table A> The List of Topics from the News Articles that Contain "Finance" in Contents

Topics	Keywords
Economy/Finance	economy, exports, growth rate, economy, interest rate, growth, Korea, Bank of Korea, forecast, governor, fiscal policy, slowdown
Employment/Welfare	union, payment, child, wage, allowance, income, support, childbirth, household, predecessor, in-house, labor and management, children, workers, childcare
Public Institutions	fund, public institution, culture, audit, foundation, institution, public enterprise, operation, auditor, financial stabilization, accumulation, contribution, art, operation, establishment
Education/Welfare	education, school, Office of Education, Nuri Course, Provincial Office of Education, Superintendent of Education, education finance, students, Ministry of Education, grants, free meals, budget, kindergarten, local education finance grants, local, university, limited, Ministry of Education, tuition, support, project
Education/Policy	private high school, student, education, school, selection, high school, general high school, selection, SAT, application
Transportation/ Infrastructure	project, litigation, railway, construction, Ministry of Land, road, four major rivers, private investment, private investment project, judgment, section, construction, promotion, toll fee, KORAIL, bus, city bus, route, completion system, fees, subsidies, public transportation
National/Financial	government, tax revenue, deficit, budget, expenditure, income, increase, national debt, income and expenditure, contrast, finance, forecast, fiscal soundness, budget proposal, supplementary, economy, government, organization, execution, finance, Ministry of Strategy and Finance, expansion, Minister, investment
International/Economy	China, Korea, Governor, conference, Minister, Ministry of Finance, economy, finance, BOK, Asia, global, loan, trade, world, attendance, fiscal cliff, United States, negotiation, index, decline, dollar, stock market, rise, Obama, settlement, record, republican
Company/Industrial	industry, support, company, fostering, small and medium-sized business, start-up, investment, innovation, field, region, construction, creation, competitiveness, attraction, industrial complex, company, Japan, energy, material, parts, regulation, investment
Ministry of Finance/Information	reorganization, Congressman, leakage, information, impression, Jaecheol Shim, Hankyung, Data, Ministry of Strategy and Finance, Office of Congress, Local Council, Auditor, Financial Information Service, D'Brain, Korea Financial Information Service, Ministry of Strategy and Finance, Ministry of Finance, planning, personnel, policy, general affairs
Health/Welfare	health insurance, hospital, insurance, medical care, health care, patient, premium, health, coverage, medical expenses, deficit, medical treatment, medical expenses, nursing hospital, treatment
Fraudulent Supply and Demand	property, leakage, collection, payment, report, illegal supply, welfare, subsidy, detection, confirmation, salary, recipient, amount, payment, secretary hospital
North Korea	North Korea, North and South, Korean Peninsula, unification, security, peace, diplomacy, provocation, missile
Incident/accident	financial application, prosecution, court, case, non-prosecution, dismissal, disposition, charge, prosecution, lawmaker, prosecutor, accusation, violation, accusation, objection, auditor, audit, corruption, investigation, detection, police, statistics, investigation, public official, allegation, Unfairness, embezzlement, prosecution, crime, prosecution, president, prosecution, investigation, suspicion, government, police, people, NIS
Safety/facility	safety, installation, Gwangju city, facilities, road, firefighting, bicycle, rearrangement, construction, Nam-gu, area, accident, city, manual, Gwangju

<Table A> The List of Topics from the News Articles that Contain "Finance" in Contents (Cont.)

Topics	Keywords
Safety/Disaster	Buk-gu, Daegu, Asan-si, Gyeongbuk-do, Busan, Gyeongbuk, Guje Station, Gangwon-do, Kim Jae-jeong, Das, Hoengseong-gun, Jeongseon-gun, Jeonju, metropolitan
Pension/Welfare	pension, national pension, public employee pension, basic pension, estimate, public official, reform, reform proposal
Job/Employment	jobs, employment, income, minimum wage, support, small and medium-size enterprises (SMEs), wages, creation
Disaster	disaster, reorganization, river, damage, safety, Sewol ferry, MERS, road, park, maintenance, typhoon, accident, facility
Finance/Evaluation	project, evaluation, ministries, insufficient, Ministry of Strategy and Finance, management, performance, Ministry of Finance, fine dust, budget, Iksan City, financial projects, civil petition, results, evaluation, selection, excellent, finance, grade, local government, soundness, local government, field, reduction, festival, incentive, national, operation, efficiency
Politics	Member of Parliament, Representative, Democratic Party, Chairman, Saenuri Party, House Representative, Free Korea Party, general election, National Assembly Member, election, presidential election, running, National Assembly, Bae Jae-Jung, election, elect, Ministry of Finance, Transition Committee, Minister, mobile, fuel tax, cut, work report, price, countermeasures, family site, official, ministries
Tax	taxation, reduction and exemption, special order, possession tax, real estate, tax, deduction, tax, tax rate, corporate tax, income tax, tax, non-taxation, reorganization, taxation
Local/Autonomous	committee, ordinance, integration, participation, committee, ordinance, composition, residents, opinion, operation, presentation, Chairperson, Cheongju City, Budget, hosting, local government, Suwon, local government, local finance, Gyeonggi Province, Ministry of Government Administration and Home Affairs, Seongnam City
Local/Financial	financial independence, local government, average, nationwide, autonomous district, local tax, financial independence, Jeonnam, collection, arrears, Seoul, income, ratio, Gwangju, share, local government, local tax, local finance, acquisition tax, burden, local, government, local consumption tax, subsidy, finance, reduction, increase, local government, conservation, Gyeonggi-do, Gyeongnam-do, brand, Gyeongnam, answer, Changwon-si, company, investigation, store, Jinju-si, Changwon, Yangsan-si, Miryang-si, Hamyang-gun, sales, financial crisis, Local debt, debt, repayment, issuance
Local/Business	citizen, village, support, education, participation, activity, program, area, held, business, operation, center, workshop, offer, event

<Table B> Quarterly Polarity Score of News Articles about "Job/employment"

Date	Negative	Positive	Neutral	Pos_rate
2010Q1	348	197	13	0.353047
2010Q2	163	163	5	0.492447
2010Q3	222	117	3	0.342105
2010Q4	232	165	13	0.402439
2011Q1	77	94	2	0.543353
2011Q2	220	97	5	0.301242
2011Q3	319	133	5	0.291028
2011Q4	129	107	8	0.438525
2012Q1	554	417	12	0.424212
2012Q2	214	61	0	0.221818
2012Q3	82	48	3	0.360902
2012Q4	179	100	2	0.355872
2013Q1	201	123	3	0.376147
2013Q2	66	81	3	0.54
2013Q3	111	107	1	0.488584
2013Q4	152	109	0	0.417625
2014Q1	262	110	4	0.292553
2014Q2	110	97	3	0.461905
2014Q3	208	141	4	0.399433
2014Q4	47	80	2	0.620155
2015Q1	109	94	1	0.460784
2015Q2	327	220	2	0.400729
2015Q3	134	105	1	0.4375
2015Q4	67	37	1	0.352381
2016Q1	87	87	1	0.497143
2016Q2	554	320	18	0.358744
2016Q3	180	94	7	0.33452
2016Q4	113	143	2	0.554264
2017Q1	93	130	2	0.577778
2017Q2	234	198	3	0.455172
2017Q3	377	587	10	0.602669
2017Q4	315	180	0	0.363636
2018Q1	592	592	6	0.497479
2018Q2	1081	888	4	0.450076
2018Q3	1542	1196	1	0.436656
2018Q4	491	367	0	0.427739
2019Q1	629	489	2	0.436607
2019Q2	691	537	1	0.436941
2019Q3	404	456	0	0.530233
2019Q4	944	1077	0	0.532905

◆ About the Authors ◆

Hansol Lee

Hansol Lee is a Teaching Assistant Professor in the department of e-Business at School of Business, Ajou University. He received his Ph.D. in Business Administration (Management Science & Operations Research) from Ajou University in 2022. His research interests are Business Analytics, Optimization and Simulation.

Juyoung Kang

Juyoung Kang is a Full Professor in the department of e-Business at School of Business, Ajou University. She received her Ph.D. in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2005. She has more than 80 refereed publications in academic journals and has developed Intelligent Systems and E-Commerce applications with various industrial partners. She has also served as a Editor-in-Chief and an editorial board member of serveral academic journals. Her research interests are in the fields of Text Mining, Big Data, ERP, Cloud Computing, and Intelligent Systems.

Sangun Park

Sangun Park is a Full Professor in the department of management information systems, Kyonggi University. He received his Ph.D in Management Engineering from the Graduate School of Management at the Korea Advanced Institute of Science and Technology (KAIST) in 2006. His research interests include deep learning, machine learning, artificial intelligence and big data analysis. He serves as an editorial board member and reviewer for a number of conferences and academic journals.

Submitted: February 25, 2022; 1st Revision: July 26, 2022; Accepted: August 16, 2022