Journal of Economic Theory and Econometrics, Vol. 35, No. 4, Dec. 2024, 87–112

Sentiment Matters in Stock Market: Construction of Sentiment Index Using Machine Learning*

Seiwan Kim† YooJeong Choi‡ Jisu Hwang Jeon§ Yanxin Lu¶

Abstract This study employs machine learning to analyze news article sentiment, developing a stock market sentiment index (SSI) based on this analysis. By examining the textual data from news articles, which constitute unstructured data, we aimed to capture the prevailing sentiments among market participants across the financial market. Specifically, this study utilizes The BERT model to decipher the psychological sentiment embedded in the articles through its contextualized understanding of the tone and language patterns. The variables tested included the risk aversion estimate, calculated using the VKOSPI and Bekaert's method for assessing risk aversion. The empirical analysis involving the SSI, VKOSPI, and risk aversion reveals a significant negative impact of SSI on VKOSPI and risk aversion. We further find that the news sentiment index (NSI) and SSI simultaneously exhibit a converging trend.

Keywords Machine learning, stock market sentiment index, risk aversion.

JEL Classification G12, G41.

*We are deeply grateful to the two anonymous reviewers for their thoughtful comments and suggestions. This work was supported by the research fund (Grant # 2020-1010).

†Department of Economics, Ewha Womans University, Seoul, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, Republic of Korea 03760. E-mail: swan@ewha.ac.kr.

‡Shinhan Securities Co. Ltd, 96 Uisadang-daero, Yeongdeungpo-gu, Seoul, Republic of Korea 07321. E-mail: yjchoi.econ@ewha.ac.kr.

§Department of Economics, Cornell University, 616 Thurston Ave. Ithaca, NY 14853, US. E-mail: jisu.hw678@gmail.com.

¶Corresponding author. Department of Economics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, Republic of Korea 03760. E-mail: annie.lu.0630@gmail.com.

Received November 5, 2024, Revised December 26, 2020, Accepted December 30, 2020

1. INTRODUCTION

In this study, we construct Korean stock market investors' aggregate sentiment index (SSI) by employing machine learning and news article big data. Noelle-Neumann (1974) discovered that personal opinions are likely to become mainstream after a certain degree of exposure, and these opinions can subsequently predict societal and political developments. Thus, the assessment of economic trends by private economic entities significantly influences market dynamics. The Bank of Korea compiles and disseminates the Business Survey Index and Consumer Survey Index to quantitatively articulate the perceptions of participants in economic activities on the economic climate. Since 2012, to incorporate a broader spectrum of private economic sentiments, key components from these indices were selectively utilized to construct the Economic Sentiment Index through a weighted average of standardized indices, reflecting higher economic correlations. However, the monthly frequency of these surveys poses a challenge in capturing the immediate economic mood, thereby limiting their responsiveness to rapid psychological shifts in stock market.

An alternative approach to gauge market sentiment involves estimating market risk aversion, as delineated by Bekaert *et al.* (2013). This methodology utilizes the S&P 500 index and its derivative, the volatility index (VIX), to distinguish between stock market uncertainty and risk aversion. The advantage of this method lies in its ability to estimate market risk aversion promptly on trading days. Nevertheless, it does not account for the limitations on emotional index generation during weekends or holidays.

Recent advancements in natural language processing (NLP) technology, particularly transformer-based models like BERT (Devlin, 2018)), have enabled the analysis of unstructured text data, leading to a significant increase in academic interest. Text data analysis has been utilized in space design to understand public perceptions of events occurring in different spaces. By examining these perceptions,Park and Park (2021) proposed future directions for public design to address suicide prevention in public places. Text mining has been used in pedagogy to identify key themes relevant to first-year students. Yoon and Shin (2021) found that keywords such as "class," "idea," "school," "friend," and "professor" not only appeared frequently but also had high centrality in the semantic network, indicating their importance to first-year students. However, the study also revealed that these first-year students perceive limitations regarding classes and professors, highlighting areas for potential improvement in the academic environment. In business administration, text data analysis has been applied to examine changes in consumer perceptions of food delivery platforms before and after

the COVID-19 outbreak. Jeong and Choi (2021)aimed to provide useful information for the restaurant industry's delivery service-related marketing strategies while identifying stakeholders' evolving needs in food delivery.

This study proposes the development of a stock market sentiment index (SSI) tailored to Korea's stock market conditions, using both traditional machine learning methods and transformer-based techniques like BERT to discern economic impacts and explore correlations with existing market indicators. Transformer models, such as BERT, are particularly effective in capturing the contextual nuances of sentiment within financial news articles, enabling a more accurate and detailed sentiment classification. This research will re-assess current indicators to comprehend market sentiments, review various machine learning methodologies, and select the most suitable ones for constructing the index. Moreover, since machine learning techniques are used to analyze previously written articles, the human resources required can be minimized by automating the process of understanding the financial market's sentiment.

This paper is structured as follows: Section 2 reviews prior studies that utilized machine learning for economic sentiment analysis. Next, Section 3 describes the data utilized in this research. In Section 4, a sentiment analysis is performed, and an index is developed. Section 5 assesses the robustness of SSI with other market indices, and Section 6 concludes while discussing the implications of the study.

2. LITERATURE REVIEW

Research into deciphering the mood of financial markets via textual data has been advancing. A prominent example is the daily news sentiment index (NSI) developed by Shapiro *et al.* (2022) at the San Francisco Federal Reserve. This index measures sentiment by assigning emotional values to individual words and computing their average. It utilizes articles from LexisNexis covering the U.S. economy from 1980 to 2015, specifically selecting those longer than 200 words and containing quotations to capture richer emotional content. The emotional value of articles was constructed using data from LexisNexis, covering articles from 16 media outlets between 1980 and 2015. Only articles focused on the U.S. economy and containing more than 200 words were selected. Additionally, articles had to include quotations (e.g., said, says, told, stated, written, reported) that were believed to contain more emotional content.

The emotional value of each article was calculated as the difference between the ratio of positive words $(+1)$ and negative words (-1). Existing emotional

Word	Sentimental Value
fear	-2.5608
feared	-1.9131
fearful	-1.8535
fearing	-2.0886
fears	-2.5565

Table 1: EXAMPLE OF THE SAN FRANCISCO FEDERAL RESERVE'S SENTIMENT DICTIONARY. The weights assigned to the same words vary in the San Francisco Federal Reserve's Sentiment Dictionary. There are also common words such as do and have etc.

dictionaries, such as VADER, were referenced to create a word-by-class matrix, which calculated the co-occurrence rate of the analyzed article and words. This matrix was then refined using the pointwise mutual information technique to determine whether each word conveyed a positive, negative, or neutral sentiment within the context of the sentence.

Subsequently, the tone of each article was re-analyzed using the enhanced emotional dictionary, which combined the words and their assigned emotional values determined through this process.

Despite its advances, the word-based sentiment analysis method encounters limitations. The assignment of values to a broad lexicon may fail to capture the nuanced meanings that different word forms convey. This issue is highlighted by the varying sentiment values assigned to the word "fear" and its derivatives across various contexts, as illustrated in Table 1.

The research work by Kim *et al.* (2021) investigates the impact of psychological factors on consumer decisions in Korean non-life insurance markets, using sentiment analysis of news articles containing the keyword "insurance." Their findings establish a Granger causality between long-term and accidental injury insurance premiums.

There have been efforts to develop an Economic Sentiment Index (ESI) by utilizing a sentiment lexicon categorized into positive, negative, and neutral sentiments, based on unigram analysis (Song and Shin, 2017). This approach involved extracting vocabulary from news articles and constructing a financial sentiment lexicon, where positive, neutral, and negative words were assigned values of $1, 0$, and -1, respectively. The ESI was then calculated by averaging these values across the total number of words. A predictive model was subsequently

developed to forecast economic indicators, such as the retail sales index, using the calculated ESI. The findings indicated that the model demonstrated enhanced predictive accuracy and goodness of fit when compared to models utilizing the Consumer Sentiment Index as a predictor.

Similarly, the Bank of Korea has introduced a comparable index as an experimental statistic. Specifically, the News Sentiment Index (NSI) was designed to quantify economic sentiment as reflected in economic news articles. Since February 2022, the NSI has been published on a weekly basis as an experimental statistic via the Bank of Korea's Economic Statistics System (ECOS). The index is compiled through the collection of economic news texts from internet portals via web crawling techniques, and these texts are analyzed using advanced natural language processing (NLP) models (Seo and Lee, 2022). In contrast to the San Francisco Federal Reserve, which constructed its index based on a sentiment lexicon, the Bank of Korea's NSI focuses on sentence-level analysis rather than individual words. The index employs a transformer model based on artificial neural networks to classify sentences into positive, negative, and neutral categories. A transformer model is an artificial neural network architecture that features a multi-head attention mechanism. This mechanism assigns higher weights to specific values within the input vectors that necessitate concentrated learning in sequential prediction models.

The NSI is generated through the daily collection of news articles, wherein sentences are processed according to the methodology described above and subsequently classified into 'domestic-positive sentences' and 'domestic-negative sentences.' The index is then compiled by counting the number of sentences in each category. However, this counting method is not without limitations, as sentence segmentation may vary depending on the journalist's writing style. For instance, negative phrases may be grouped into a single sentence using conjunctions, while positive phrases might be split into separate sentences, potentially leading to distortions.

3. DATA

News writing not only reflects the journalist's thoughts but is also influenced by social opinions (Lang and Lang, 1983). Hence, this paper employs a method of analyzing news to extract a stock market sentiment index.

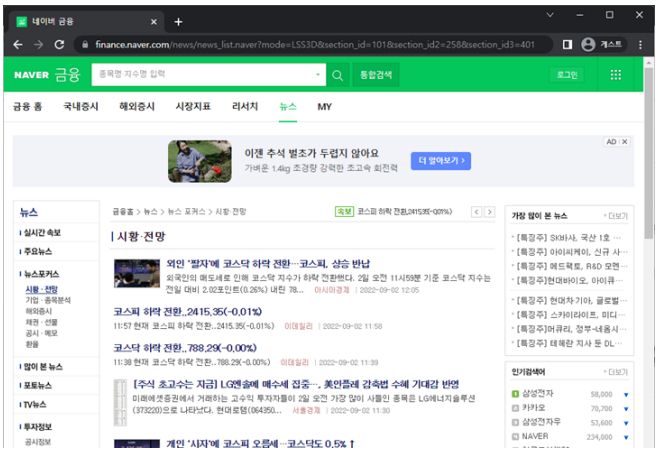

Figure 1: NAVER FINANCIAL MARKET FORECAST. The picture displays a specific webpage captured by the news crawling. It starts from Naver's main page, navigates to the Finance section, and proceeds to the News column.

3.1. UNSTRUCTURED DATA

This study employs web scraping to extract textual data from news articles for the construction of SSI. Web scraping is a method used to extract data directly from websites via software, storing the data in a structured format. Although web scraping is often used interchangeably with web crawling, they are distinct processes. Web crawling specifically refers to accessing a given URL, discovering related URLs, and continuously identifying, categorizing, and storing hyperlinks from these URLs. This process allows the crawler to traverse multiple web pages, create an index of data locations, and store the information in a database.

This study focused on collecting unique news articles, excluding those presented in tables or images, to preserve data integrity and avoid skewing the index values. The initial publication time of each article was meticulously recorded, with subsequent redundant publications excluded to ensure the accuracy of the SSI.

The "Market Conditions and Forecasts" section of Naver Financials, known for its substantial market presence and detailed coverage of current and projected market conditions, was selected for data collection. The site is shown in Figure 1 above. This choice aligns with the study's objective to capture the prevailing mood among market participants. The focus of this paper is on the aggregate market sentiment as perceived by participants in the financial market. Therefore, the market outlook and forecast sections were chosen. These sections provide

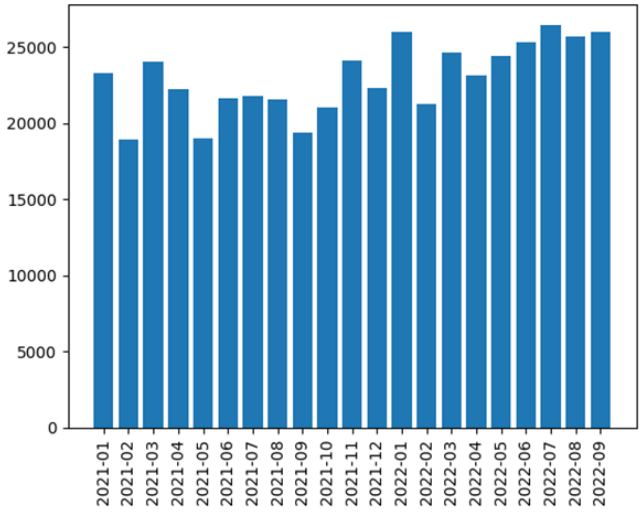

Figure 2: NUMBER OF SENTENCES COLLECTED PER MONTH. Figure 2 illustrates the monthly counts of the news articles captured from January 2021 to September 2022, generally between 20,000 and 25,000.

insights not only into the current state of the financial market but also into the anticipated future financial conditions as projected at the present time.

Data collected through web scraping may include articles containing only tables or photos and duplicate articles or content. These articles can undermine the integrity of the data and distort the value of the index due to the lack of additional information. Such articles were removed to ensure the accuracy of the SSI. Therefore, the initial publication date was considered when the news was first released, and any duplicate data published subsequently were excluded from the analysis.

From January 1, 2021, to September 30, 2022, a total of 35,668 articles were scraped from Naver Financials on business days only, enabling a direct comparison with existing financial indicators in Figure 2 below. This collection was then segmented into 485,234 sentences to minimize subjective interpretations that could arise from analyzing full articles.

To address seasonal or period-specific biases in the news data, 3,093 sentences were randomly selected from the total dataset and categorized by ten economics graduate students as positive, neutral, or negative. After initial preprocessing, a portion of the content was manually labeled as positive, negative, or neutral. This rigorous sentence-level analysis identified 1,171 sentences as

positive, 1,080 as neutral, and 842 as negative, forming a robust foundation for a nuanced SSI. Given that the interpretation of emotions is inherently influenced by individual perspectives, it is inevitable that manually labeled emotions reflect certain subjective biases (Seo *et al.*, 2024).

These 3093 labeled sentences have been divided into two parts with a ratio of 80:20. 80% is the training set and 20% is the test set. They are used equally for training each model.

3.2. STRUCTURED DATA

3.2.1. Volatility Index: VKOSPI

The volatility index, commonly referred to as the "fear index," is crucial in financial markets as it is generally inversely correlated with the underlying asset, such as a stock index. Widely utilized by both academic circles and industry practitioners, this index is a significant indicator and a major source of data for assessing newly constructed SSI's performances . In the context of the Korean market, after a two-year development phase, the Korea Exchange (KRX) introduced a volatility index, known as VKOSPI, tailored to the characteristics of the Korean financial environment., it has been calculated and published since April 13, 2009, providing a daily measure of volatility based on the KOSPI200 index data (Choi and Han, 2009).

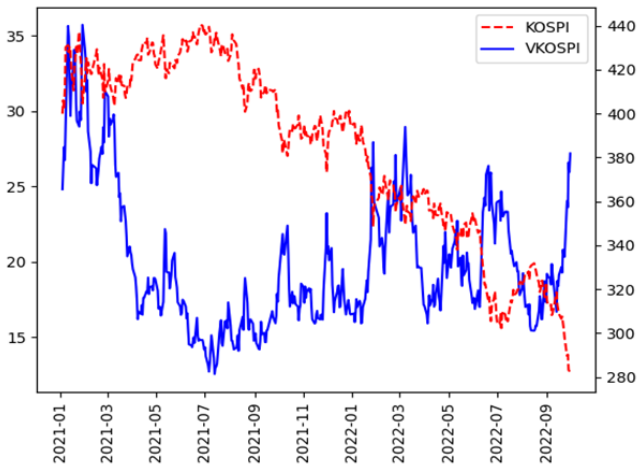

Figure 3: TRENDS OF KOSPI AND VKOSPI. The VKOSPI Index and the KOSPI200 Index are sourced from the Korea Exchange (KRX) Market Data database.

Figure 3 shows the trend of daily VKOSPI index along with KOSPI 200's movement on daily basis. The graph illustrates that the volatility and stock index are inversely proportional.

3.2.2. Risk Aversion Estimate

The volatility index is frequently interpreted as a measure of market fear. Bekaert *et al.* (2013) extended this interpretation by disentangling market uncertainty and risk aversion from the volatility index, allowing for the estimation of the market's level of risk aversion.

The VIX represents the implied expected volatility of S&P 500 index options. This measure, known as implied volatility or risk-neutral volatility, is distinct from expected (or physical) volatility. Physical volatility refers to the expected volatility derived from probabilities observed in the real world. In contrast, risk-neutral volatility is based on a risk-neutral framework, where it is derived from risk-neutral probabilities, ensuring that the expected return on options equals the risk-free interest rate.

Both the VKOSPI and risk aversion data were estimated for the period from January 1, 2021, to September 30, 2022, consistent with the news article collection period. The daily variance of the KOSPI was calculated, with the daily realized variance (RVAR) derived from the log differences between the intraday high and low prices. The reason for considering volatility levels through both VKOSPI and RVAR is to measure the risk aversion tendencies of economic agents using the VKOSPI index, which reflects volatility levels under risk-neutral conditions (Bekaert *et al.*, 2013).

Risk-neutral valuation is a principle in which investors assume a risk-neutral stance when evaluating derivatives. In this framework, attitudes toward risk are not considered important when valuing an option based on the price of an underlying asset. By assuming a risk-neutral market, it is possible to calculate accurate prices not only in the risk-neutral scenario but in all possible scenarios (Hull and Basu, 2016).

The risk-neutral probability is the probability that ensures the expected return equals the risk-free interest rate. When there is a negative rate of return or high volatility, this probability assigns a higher weight than the actual probability. Consequently, the VKOSPI index, calculated based on the risk-neutral probability, tends to be larger than the RVAR, which is a conditional variance value derived from the actual probability. This characteristic of risk-neutral probability means that in scenarios of negative returns or high volatility, it assigns greater weight compared to the actual probability.

The difference between the VKOSPI index and the conditional variance value (RVAR) is indicative of the risk aversion tendency. This disparity reflects the risk levels perceived by market participants, which are embedded in the VKOSPI index. Thus, the VKOSPI index was divided into two components, i.e., risk aversion and pure uncertainty, and used for empirical work in this study.

Uncertainty Regression analysis was conducted to estimate conditional variance following the methodology of (Bekaert *et al.*, 2013).

$$\mathrm{RVAR}_t = \beta_0 + \beta_1 \mathrm{VKOSPI}_{t-1}^2 + \beta_2 \mathrm{RVAR}_{t-1} + e_t. \tag{1}$$

The variable meanings in the regression equation are presented as follows:

RVAR_t: This represents the conditional variance at time t, which serves as a measure of market risk. It often reflects volatility or the degree of uncertainty in the market.

β_0: This is the intercept term, indicating the estimated value of RVAR when other variables are zero.

VKOSPI_{t-1}^2: This is the squared value of VKOSPI from the previous period $(t-1)$. VKOSPI usually refers to the volatility index of the Korean stock market (similar to the VIX in the US). Squaring it is common in modeling conditional variance, as it captures the variations in the risk premium.

RVAR_{t-1}: This is the lagged conditional variance (RVAR) from the previous period. Including the lagged term allows the model to capture the autocorrelation in risk over time, acknowledging that volatility tends to persist.

e_t: Error term, representing random disturbances or the unexplained portion in the model.

In this analysis, RVAR was calculated using the logarithmic difference between the highest and lowest prices of the day.

The estimate of uncertainty UC_t was obtained as a fitted value explained solely by the two explanatory variables, excluding the residuals in (1).

$$\mathrm{UC}_t = \widehat{\mathrm{RVAR}_t} = \beta_0' + \beta_1' \mathrm{VKOSPI}_{t-1}^2 + \beta_2' \mathrm{RVAR}_{t-1}. \tag{2}$$

Consequently, the final estimate of uncertainty was derived using (2).

Figure 4 presents a trend graph illustrating VKOSPI and uncertainty. To address the inconsistency in their units, both variables have been standardized.

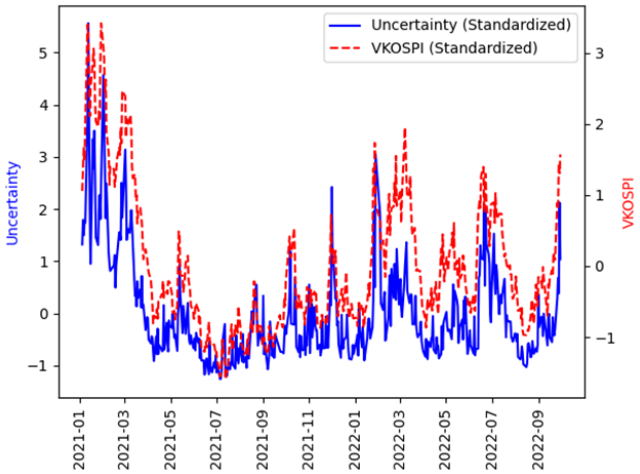

Figure 4: TRENDS OF VKOSPI AND UNCERTAINTY. The VKOSPI index is collected from the Korea Exchange (KRX) market data database and the uncertainty is calculated using (2).

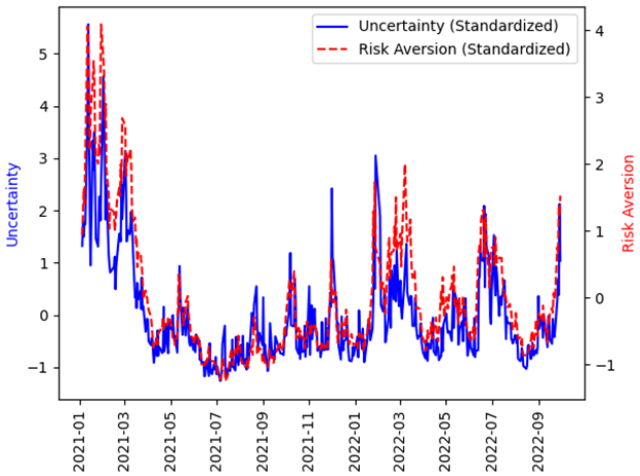

Figure 5: TRENDS OF UNCERTAINTY AND RISK AVERSION. The Risk Aversion shown in the figure is calculated using (3). To account for the significant differences in units, standardization was applied.

Risk Aversion The estimated level of risk aversion, RA_t, was derived by subtracting the uncertainty calculated in (2) from the VKOSPI index, which is the level of implied volatility. This is based on the assumption of risk neutrality, and only factors related to risk attitude were derived from the calculated VKOSPI index, excluding pure uncertainty. Specifically, the following was used:

$$\mathrm{RA}_t = \mathrm{VKOSPI}_t - \widehat{\mathrm{RVAR}_t}. \tag{3}$$

Figure 5 shows the trend of the finally derived risk aversion and uncertainty. As in Figure 4 , both variables have been standardized to address the problem of inconsistent units. It was found that they generally tended to move together.

4. CONSTRUCTION OF STOCK MARKET SENTIMENT INDEX

4.1. TEXT EMBEDDING

The transformation of text data from financial news articles into a format amenable to machine learning analysis entails a sequence of crucial preprocessing steps. This process involves converting text into numerical forms recognizable by computational models, incorporating several sub-procedures, such as tokenization and the removal of non-essential textual elements.

4.1.1. Preprocessing

The initial phase of preprocessing involves breaking down text sentences into smaller, meaningful units called morphemes, which are the smallest grammatical units in a language. For example, the Korean sentence "최근 부진을 겪었던 국내외 증시는 실적 시즌에 돌입하며 하락세가 진정되는 양상이다" would undergo tokenization, resulting in morphemes such as "최근/Noun," "부진/Noun," "을/Particle," and "겪었던/Verb."

4.1.2. Text Embedding Techniques

Text embedding converts textual data into numerical representations for machine learning models. The term "embedding" signifies the transformation of text into a vector space that captures semantic or contextual information. Traditional embedding techniques, such as the N-gram model, process tokens based on their frequency and sequence within a specified window size. While these approaches improve upon simple word-level embeddings by incorporating some contextual information, they are inherently limited in scope.

For instance, in a unigram model, individual words like “sluggish” and “recovery” are treated independently, potentially losing the negative connotation of their combination. A bigram model processes two words together, recognizing the phrase “sluggish recovery” as conveying negative sentiment. However, the fixed window size of N-gram models restricts their ability to capture semantic dependencies or nuances in longer sentences.

To address these limitations, the TF-IDF method is often applied in conjunction with the N-gram model. TF-IDF is a statistical technique that evaluates the relative importance of words by weighting their frequency and distinctiveness. The formula for TF-IDF is as follows:

Text embedding refers to the process of converting textual data into numerical representations for use in statistical and machine learning models. The term “embedding” signifies the transformation of text into a vector space that captures semantic or contextual information. Among traditional embedding techniques, the N-gram model is often utilized to define the number of tokens processed simultaneously. While this approach offers improvements over word-level embeddings by capturing some degree of contextual meaning, it remains limited in its scope.

For instance, in a unigram model, individual words like “sluggish” and “recovery” are treated independently, potentially leading to misinterpretation of their combined meaning. Conversely, a bigram model processes two words together and can recognize the negative sentiment in the phrase “sluggish recovery.” Despite this advantage, N-gram models are inherently constrained by their fixed window size, which prevents them from capturing dependencies or semantic nuances beyond the specified range. This limitation can result in a loss of critical contextual information, especially in longer or more complex sentences.

In this study, the TF-IDF method was applied in conjunction with the N-gram model to embed tokenized words. TF-IDF is a statistical technique that evaluates the relative importance of words by weighting their frequency and distinctiveness. The formula for TF-IDF is as follows:

The term frequency-inverse document frequency (TF-IDF) is defined as follows:

$$\text{TF-IDF}(w) = \text{TF}(w) \times \log \frac{N}{\text{DF}(w)}, \tag{4}$$

where TF: term frequency, indicating how often a word w appears in a specific document; DF: document frequency, denoting the number of documents containing the word w; N: total number of documents; IDF: inverse document frequency, measuring the rarity of w across the corpus. A higher IDF value suggests that the

word appears infrequently and is therefore more unique to specific contexts, enhancing its predictive ability for topic identification.

Despite its utility, TF-IDF also has limitations. It relies solely on word frequency statistics, ignoring word order and deep semantic relationships. Moreover, its term-based vectorization often results in high-dimensional feature spaces, increasing computational complexity and resource requirements. Meanwhile, high-dimensional feature spaces can impact the effectiveness of certain machine learning algorithms, particularly non-linear models (Kowsari *et al.*, 2019).

To overcome these challenges, advanced techniques like BERT (Bidirectional Encoder Representations from Transformers) leverage deep contextualized embeddings. BERT considers the full sentence context bidirectionally, capturing richer semantic and syntactic features. BERT defines the embedding process as:

The input embedding is defined as:

$$E_{\text{input}}(w) = E_{\text{token}}(w) + E_{\text{position}}(w) + E_{\text{segment}}(w),$$

where $E_{\text{token}}(w)$: token embedding: a static embedding derived from a pre-trained vocabulary, representing a token's semantic information independent of context; $E_{\text{position}}(w)$: positional embedding: encodes the token's position within the input sequence, ensuring awareness of token order; $E_{\text{segment}}(w)$: segment embedding: differentiated tokens belonging to distinct input segments (e.g., sentence pairs in tasks like question answering).

Finally, the input embeddings are fed into a transformer model to compute context-aware representations:

$$E_{\text{BERT}}(w) = \text{Transformer}([E_{\text{token}}(w), E_{\text{position}}(w), E_{\text{segment}}(w)]).$$

By stacking multiple transformer layers, BERT generates dynamic and semantically rich embeddings, addressing the shortcomings of traditional models like TF-IDF and N-grams.

4.2. IMBALANCED DATA

The dataset used in this study comprises 3,093 labeled sentences divided into three sentiment classes: positive, neutral, and negative. The distribution reveals an imbalance, with 1,171 sentences (37.8%) labeled as positive, 1,080 sentences (34.9%) as neutral, and 842 sentences (27.2%) as negative. This imbalance poses a challenge for training machine learning models, as they may tend to favor the majority class, resulting in reduced performance for minority classes. To address

this, class weights were incorporated into the loss function during training, ensuring that errors on minority classes were penalized more heavily. Additionally, evaluation metrics such as F1-score, precision, and recall were used to better assess the model's performance on all classes, especially the underrepresented negative sentiment.

To handle the inherent class imbalance in the dataset, a weighted cross-entropy loss function was employed in the BERT model. The weights were computed as the inverse class frequency, ensuring that the model penalizes errors on minority classes more heavily. This approach directly mitigates the impact of imbalanced data during training by emphasizing underrepresented classes.

$$w_c = \frac{N}{|C| \times n_c},$$

where w_c is the weight for class c; N is the total number of samples; $|C|$ is the total number of classes; and n_c is the count of samples in class c. The computed weights are then integrated into the cross-entropy loss function to penalize misclassifications in underrepresented classes more heavily.

In the case of the traditional models, such as XGBoost, we have utilized its built-in capability to handle imbalanced datasets effectively (Chen and Guestrin, 2016). Adjustments to the objective function balanced the weights assigned to each class during training. Therefore, we did not specifically process the imbalanced data when comparing with the traditional model.

In addition, the model's performance is evaluated using metrics such as F1-score, precision, and recall, which provide a more balanced assessment of effectiveness across all classes, addressing potential bias caused by class imbalance.

4.3. SENTIMENT CLASSIFICATION MODELS AND SSI

The following five traditional machine learning methods were used for the classification of news sentiment: random forest, support vector machine (SVM), light gradient-boosting machine (LightGBM), extreme gradient boosting (XGBoost), and logistic regression. In addition to these methods, this study also incorporated BERT (Bidirectional Encoder Representations from Transformers) to leverage its ability to capture rich contextual information and its effectiveness in natural language understanding. Among these methods, BERT demonstrated superior accuracy. In the followings, we briefly describe the features of each method.

Traditional Machine Learning Methods Random forest is an ensemble learning technique that constructs numerous small decision trees during training. It selects random subsets of features for each tree and uses the mode of the class labels (most frequent value) predicted by individual trees for the final prediction. Known for its effectiveness in handling large datasets, Random forest helps prevent overfitting by allowing for the adjustment of the depth of decision trees (Breiman, 2001).

SVM is a robust classifier that determines the best hyperplane to maximize the margin between different classes. It is particularly beneficial for high-dimensional spaces. SVM works well with limited data relative to the number of dimensions and is effective in high-dimensional spaces (Hearst *et al.*, 1998).

LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It grows trees leaf-wise (vertically) rather than level-wise (horizontally), allowing for faster learning and better efficiency with large datasets. It is efficient with large datasets and capable of faster execution due to its leaf-wise growth strategy (Ke *et al.*, 2017).

XGBoost is an optimized distributed gradient-boosting library. It enhances the performance and speed of gradient-boosting models by using several weak decision trees and learning through parallel processing. Known for its execution speed and performance, XGBoost can handle large datasets effectively and is versatile across various types of predictive modeling (Chen and Guestrin, 2016).

Logistic regression is a statistical model that estimates the probability of a binary outcome based on independent variables. It uses the logistic function to model a binary dependent variable, making it suitable for binary classification problems. The key advantage of logistic regression is its ability to provide the significance of individual predictors, which is useful for understanding the impact of each variable. Coefficients in logistic regression express the change in the log odds of the dependent variable for a one-unit change in the predictor variable (Kleinbaum *et al.*, 2002). Logistic regression demonstrated competitive accuracy among traditional models in this study, highlighting its simplicity and interpretability for sentiment classification.

4.4. BERT MODEL WITH FINANCIAL DICTIONARY FINE-TUNING

BERT is a deep learning-based model that leverages the transformer architecture to generate contextualized word embeddings, enabling it to capture complex semantic and syntactic relationships within text. Unlike traditional methods, BERT processes sentences bidirectionally, allowing it to account for the full context of a word's usage in a sentence.

Model \ Grade	Accuracy	Precision	Recall	F1-score
BERT	0.8834	0.9013	0.8974	0.8994
Logistic	0.7643	0.7264	0.9530	0.8244
RF	0.7593	0.7290	0.9316	0.8180
SVM	0.7568	0.7193	0.9529	0.8198
LGBM	0.7171	0.6807	0.9658	0.7986
XGB	0.7246	0.7412	0.8077	0.7730

Table 2: PERFORMANCE COMPARISON BY MACHINE LEARNING ALGORITHM. The accuracy defined as the proportion of correctly predicted samples out of the total samples in the test set is used as the primary evaluation metric, particularly in cases where the class distribution is relatively balanced.

In this study, a pre-trained BERT model was fine-tuned on a financial news dataset to enhance its domain-specific understanding. Additionally, a financial dictionary was integrated during the fine-tuning process to further tailor the model to the unique terminologies and expressions commonly used in financial contexts. The incorporation of the financial dictionary allowed BERT to better interpret domain-specific terms and improved its ability to distinguish subtle sentiment cues in financial news articles.

The following outlines the structure of the BERT-based sentiment classification model:

$$E_{\text{BERT}} = \text{Transformer}(E_{\text{input}}). \tag{5}$$

The output embedding from BERT was passed through a softmax layer to predict the sentiment class.

4.5. PERFORMANCE COMPARISON

While traditional machine learning methods such as logistic regression and XGBoost achieved competitive performance in sentiment classification, BERT outperformed these models significantly. Its ability to model contextual dependencies and leverage domain-specific knowledge through fine-tuning made it particularly effective for this task. Table 2 summarizes the performance metrics, demonstrating BERT's superiority in terms of accuracy, precision, recall, and F1-score.

Figure 6, presents the confusion matrix for the BERT model. It is worth noting that although the number of negative samples in the training set is signif-

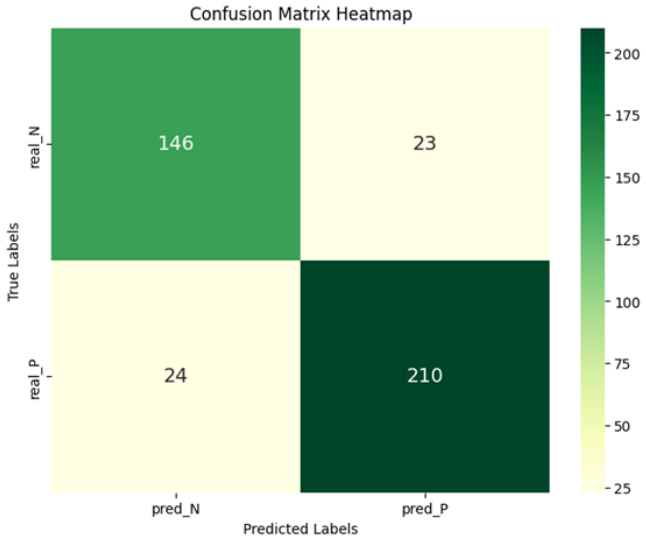

Figure 6: CONFUSION MATRIX GENERATED BY BERT. In cases of class imbalance, the confusion matrix provides a clear representation of the predictive performance for each class, preventing the overall accuracy from masking potential issues.

icantly smaller than that of positive samples, the implementation of unbalanced data processing has resulted in nearly identical prediction rates for both classes.

The confusion matrix provides an overview of the model's classification performance, showing 146 true negatives, 210 true positives, 23 false positives, and 24 false negatives. With these results, the model achieves an overall accuracy of 90%, demonstrating a strong ability to classify sentiment correctly. The precision (90%) indicates that most predictions labeled as positive were accurate, while the recall (89.7%) shows that the model successfully identified nearly all positive samples. The F1-Score of 89.85% highlights the balance between precision and recall, reflecting the model's overall reliability.

However, the error analysis reveals areas for improvement. The 23 false positives suggest that certain negative samples may contain features resembling positive sentiment, leading to misclassification. Similarly, the 24 false negatives indicate that some subtle or ambiguous expressions of positivity were missed by the model. To enhance performance, feature engineering, hyperparameter tuning, and data augmentation could be employed. Despite these limitations, the model demonstrates robust performance, particularly in identifying positive sentiment, making it a reliable tool for sentiment analysis in financial contexts.

Statistic	Value
Mean	0.0000
Standard Deviation	0.0273
Minimum	-0.1037
25%	-0.0155
Median	0.0024
75%	0.0203
Maximum	0.0575

Table 3: FINANCIAL SENTIMENT INDEX STATISTICS. The figures provide descriptive statistics of the sentiment index data, enabling a clear understanding of its distribution, trends, and characteristics.

Table 3 with a mean value of 0.00 and a standard deviation of 0.0273, reflects a predominantly neutral sentiment across the analyzed period. The narrow range between the minimum (-0.1037) and maximum (0.0575) values, coupled with the small interquartile range (0.0358), indicates limited variability in sentiment. This stability suggests a balanced reporting tone in financial news, with neither overly optimistic nor pessimistic narratives dominating. Extreme negative values, such as -0.1037, likely correspond to specific adverse events, warranting further investigation into their temporal and contextual drivers. Overall, the index provides a reliable baseline for evaluating sentiment-driven market dynamics and potential correlations with key financial.

From Figure 7, we have carefully analyzed and interpreted the observed patterns to provide meaningful insights into economic and market behaviors. For example, the pronounced fluctuations in January 2021 can be linked to significant events, such as the Kospi index exceeding 3,000 points, followed by a decline due to a 218,000 year-on-year reduction in employment and a 4% unemployment rate. The relative stability from February to May likely reflects the effects of sustained export growth and COVID-19 mitigation policies.

Furthermore, the observed downward trend from July to September aligns with economic developments, including a 5.1% increase in the national minimum wage and a 3.2% year-on-year rise in the Consumer Price Index (CPI), the highest since 2012. These findings highlight the connection between inflationary pressures and sentiment trends, with June 2022 marking a sharp decline in SSI due to escalating inflation.

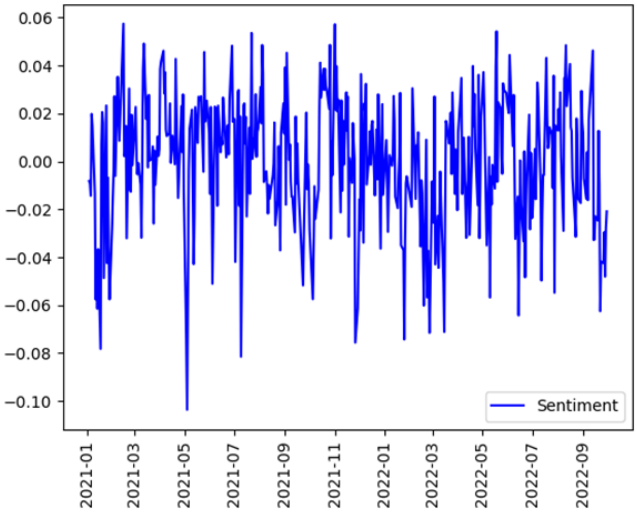

Figure 7: TRENDS IN FINANCIAL SENTIMENT INDEX. In the context of emotional fluctuations, a value greater than zero indicates a positive fluctuation, while a value less than zero signifies a negative fluctuation.

5. ROBUSTNESS OF STOCK MARKET SENTIMENT INDEX

At this stage of the analysis, the Stock Market Sentiment Index has been demonstrated as a critical indicator that reflects the overall mood of the financial market. The SSI is derived from financial news articles sourced from Naver's market status and forecast section, utilizing machine learning techniques to classify sentence-level sentiment into positive, negative, or neutral categories. The index is computed as the probability of a sentence being positive, with values ranging from -1 (highly negative sentiment) to 1 (highly positive sentiment). This makes the SSI a reliable gauge of market sentiment fluctuations.

Table 4 presents the results of the Granger causality test, evaluating the relationships between SSI and three key financial indicators: VKOSPI, risk aversion (RA), and the News Sentiment Index (NSI). The p-values indicate that SSI is a significant Granger cause of all three indicators at the 5% significance level, as detailed below:

The Granger causality test confirms that the changes in SSI Granger-cause the changes in VKOSPI, RA, and NSI. This finding highlights the predictive power of SSI in capturing shifts in market volatility, investor risk aversion, and news sentiment.

Financial Indicator	Causing Variable	p-value	Significant
VKOSPI	SSI	0.0063	Yes
Risk aversion	SSI	0.0056	Yes
NSI	SSI	0.0185	Yes

Table 4: CAUSAL TESTS BETWEEN THE STOCK SENTIMENT INDEX AND OTHER INDICATORS. The figures determine whether SSI influences the other indicators, or vice versa, using tests such as Granger causality test.

The following models the relationship between SSI and VKOSPI:

$$\text{VKOSPI}_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \text{VKOSPI}_{t-i} + \sum_{j=1}^{q} \beta_j \text{SSI}_{t-j} + e_t. \tag{6}$$

The results establish a significant negative causal relationship between SSI and VKOSPI. As financial sentiment improves (higher SSI values), the market volatility (VKOSPI) tends to decline, reflecting increased market stability and optimism. Conversely, lower SSI values, indicative of pessimistic sentiment, are associated with heightened volatility.

The relationship between SSI and RA is modeled as follows:

$$\text{RA}_t = \theta_0 + \sum_{i=1}^{p} \gamma_i \text{RA}_{t-i} + \sum_{j=1}^{q} \delta_j \text{SSI}_{t-j} + e_t. \tag{7}$$

The analysis reveals a significant negative relationship, suggesting that an increase in SSI corresponds to a decrease in RA. This indicates that positive market sentiment reduces investors' fear and risk aversion, aligning with the broader observation that heightened pessimism often accompanies deteriorating financial sentiment.

$$\text{NSI}_t = \varphi_0 + \sum_{i=1}^{p} \delta_i \text{NSI}_{t-i} + \sum_{j=1}^{q} \tau_j \text{SSI}_{t-j} + e_t. \tag{8}$$

Equation (8) explores the dynamic relationship between SSI and NSI. Unlike VKOSPI and RA, the relationship between SSI and NSI is positive. The observed correlation supports the hypothesis that improved news sentiment is strongly associated with enhanced financial sentiment.

Figure 8 illustrates the trends of the News Sentiment Index (NSI) from the Bank of Korea and the Stock Sentiment Index derived in this study. Despite

differences in data sources and computational methods, the two indices exhibit analogous trajectories, reinforcing the hypothesis that news sentiment substantially influences financial sentiment. However, the Granger causality test reveals that SSI significantly predicts changes in NSI, suggesting that financial sentiment, as captured by SSI, plays a leading role in reflecting market dynamics.

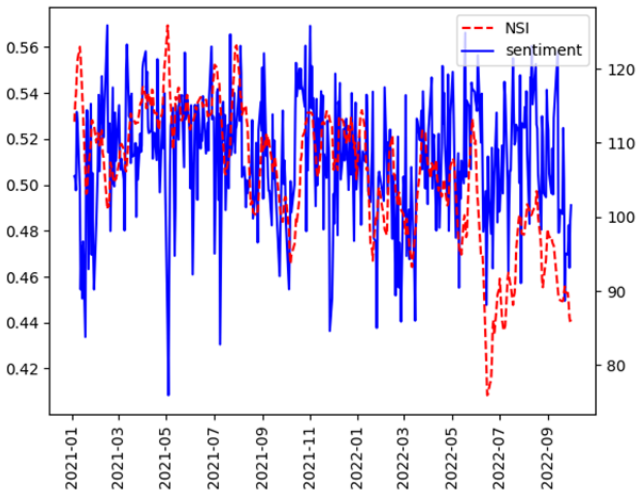

Figure 8: TRENDS WITHIN THE NEWS SENTIMENT AND SSI. Daily News Sentiment Index (NSI) is an index calculated by the Bank of Korea based on daily news.

6. CONCLUSION

This study constructs the Stock Market Sentiment Index by processing previously untapped news articles using an innovative deep learning methodology, specifically the BERT (Bidirectional Encoder Representations from Transformers) model. News data offers several advantages: it is less time-consuming and less costly to obtain compared to structured data gathered via surveys or other aggregative methods, and it provides a vast amount of real-time information. Consequently, the Stock Market Sentiment Index constructed from news data holds significant implications for financial research and practice.

Firstly, SSI demonstrates exceptional high-frequency scalability. By leveraging BERT, the model processes vast amounts of unstructured textual data in real time, enabling the generation of sentiment indices at much higher frequencies compared to traditional survey-based psychological indices. For example, while

survey-based indices are typically updated monthly, the SSI can be updated continuously, providing researchers and practitioners with timely insights that adapt easily to diverse research requirements.

Moreover, the contextualized embeddings generated by BERT allow for a deeper and more nuanced understanding of news sentiment compared to traditional methods. BERT's bidirectional architecture captures the full context of words within a sentence, significantly improving its ability to classify sentiment accurately. By fine-tuning the model on domain-specific financial data, this study ensures that the SSI reflects the unique language and sentiment patterns of financial news. Instead of analyzing simple word coefficients, as in logistic regression, the model leverages token embeddings and attention mechanisms to identify the key phrases or sentences that contribute most to the index, offering a more robust and context-aware sentiment analysis framework.

Despite the advancements brought by BERT, challenges remain regarding the interpretability of machine-generated classifications. While BERT significantly improves sentiment classification accuracy, its deep learning nature makes it less transparent compared to traditional models like logistic regression. To address this, additional interpretability techniques, such as attention visualization and SHAP (SHapley Additive exPlanations) values, are incorporated to elucidate the key factors driving the model's predictions. This enhances the reliability and usability of the SSI for academic and practical applications.

As a machine-generated sentiment index, the SSI holds potential to synergize with emerging investment strategies, particularly those driven by artificial intelligence, such as quantitative investing. By complementing other sentiment indices, it enables a more comprehensive analysis of financial markets from diverse perspectives. The contextual and real-time capabilities of BERT make the SSI a valuable tool for understanding the rapid and often unpredictable shifts in market sentiment.

In the digital era, unstructured data is continuously accumulating in real time, characterized by both vast scope and volume, and the financial market is no exception. It is becoming increasingly clear that efforts to understand the financial market through the analysis of unstructured data will gain importance. The adoption of advanced machine learning models like BERT, combined with efforts to enhance interpretability and scalability, represents a significant step forward in utilizing unstructured data to predict economic trends and inform investment strategies. Fostering the development of such experimental statistics and embracing the diversification of statistical approaches will be essential in improving our understanding and prediction of future economic trends.

REFERENCES

Bekaert, G., Hoerova, M., and Duca, M. L. (2013). "Risk, uncertainty and monetary policy," *Journal of Monetary Economics* 60, 771-788.

Breiman,L. (2001). "Random forests," *Machine learning* 45, 5-32.

Chen, T. and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, 785-794.

Choi, H. and Han, S. (2009). "Explanation and empirical analysis of the volatility index (VKOSPI)," *Key Data on KRX Market*, 43-56.

Devlin, J. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). "Support vector machines," *IEEE Intelligent Systems and Their Applications* 13, 18-28.

Hull, J. C. and Basu, S. (2016). *Options, futures, and other derivatives*, Pearson Education India.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). "Text classification algorithms: A survey," *Information* 10, 150.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye Q., and Liu, T. Y. (2017). "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems* 30, 3146-3154.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*, New York: Springer-Verlag.

Lang, G. E. and Lang, K. (1983). *The battle for public opinion: The president, the press, and the polls during Watergate*, New York: Columbia University Press.

Noelle-Neumann, E. (1974). "The spiral of silence: A theory of public opinion," *Journal of Communication* 24, 43-51.

Seo, B., Lee, Y., and Cho, H. (2024). "Measuring news sentiment of Korea using transformer," *Korean Economic Review* 40, 149-176.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). "Measuring news sentiment,"*Journal of econometrics* 228, 221-243.

김세완 · 김영민 · 구지현 · 임소연 (2021). "심리변수가 손해보험 소비자의 의사결정에 미치는 영향: 위험회피수준 및 댓글감성을 중심으로," *보험금융연구* 32, 65-93.

(Translated in English) Kim, Sei-Wan, Khu, Jeehyun, Lim, Soyeon, and Kim, Young-Min. (2021). "The effect of psychology factors on property insurance buyers' decision: New evidence from risk-aversion and online comments sentiment," *Journal of Insurance and Finance* 32, 65-93.

박변갑 · 박성룡 (2021). "텍스트마이닝 분석을 통한 공간의 의미 표출에 관한 연구 - 교량 자살 빅데이터를 중심으로," *한국공간디자인학회 논문집* 16, 181-190.

(Translated in English) Park, B., and Park, S. (2021). "A study on the expression of spatial meaning through text mining analysis - focusing on big data about suicide on the bridge," *Journal of Korea Institute of Spatial Design* 16, 181-190.

서범석 · 이영환 · 조형배 (2022). "기계학습을 이용한 뉴스심리지수 (NSI) 의 작성과 활용," *BOK 국민계정리뷰*, 68-90.

(Translated in English) Seo, B., Lee, Y., and Cho, H. (2022). "Construction and utilization of the news sentiment index (NSI) using machine learning," *BOK National Account Review*, 68-90.

송민채 · 신경식 (2017). "뉴스기사를 이용한 소비자의 경기심리지수 생성," *지능정보연구* 23, 1-27.

(Translated in English) Song, M., and Shin, K. (2017). "Construction of consumer confidence index based on sentiment analysis using news articles," *Journal of Intelligence and Information Systems* 23, 1-27.

윤태일 · 신소영 (2021). "텍스트마이닝 기법을 활용한 신입생의 대학생활 경험 분석.," *인문사회 21* 12, 279-294.

(Translated in English) Yoon, Taeil, and Shin, Soyoung. (2021). "An analysis of freshmen's college life experience through text mining," *The Journal of Humanities and Social Sciences 21* 12, 279-294.

정민경 · 최규완 (2021). "텍스트 마이닝 기법을 활용한 COVID-19 발생 이전 · 이후의 배달주문플랫폼 서비스에 대한 주요 토픽 분석," *지역산업연구* 44, 283-305.

(Translated in English) Jeong, M. Y. and Choi, K. W. (2021). "Analysis of major topics for platform services for delivery orders before and after COVID19 with the use of text mining techniques," *The Institute of Business Management* 44, 283-305.