Measuring News Sentiment of Korea Using Transformer

Beomseok Seo[*], Younghwan Lee[†], Hyungbae Cho[‡§]

Abstract

We have developed the Korean News Sentiment Index (NSI) to gauge the economic sentiment of Korea on a daily basis through the analysis of news texts gathered from the internet. Our framework utilizes cutting-edge natural language processing techniques to compute the NSI and examine keywords, offering insights into its fluctuations. To compute the NSI of Korea, we designed a sentiment classifier using transformer neural networks that effectively processes extensive news samples. By computing the NSI more frequently and immediately than official indices that rely on monthly surveys, we can identify changes in economic sentiment before official statistics are released. Moreover, the proposed framework offers keyword analysis and sector indices to clarify why economic sentiments fluctuate. Our comprehensive assessments demonstrate that the NSI is a valuable leading index and an essential tool for identifying inflection points in economic sentiment.

Keywords: news text data, natural language processing for economics, sentiment shocks.

JEL: C45, C82, E32.

1 Introduction

Economists and central bankers are interested in developing a frequently and immediately available economic indicator to gauge economic sentiment(Barsky and Sims, 2012). While

[*]Corresponding author: Economist, Department of Economic Statistics, Bank of Korea, 55, Namdaemun-ro, Jung-Gu, Seoul 04532, Korea. E-mail: bsseo@bok.or.kr

[†]Economist, Department of Economic Statistics, Bank of Korea, E-mail: yhlee@bok.or.kr

[‡]Junior Economist, Department of Economic Statistics, Bank of Korea, E-mail: hyungbae.cho@bok.or.kr

[§]This work is a substantially revised version of our working paper Seo et al. (2022). We thank Dr. Hyejung Moon, the section leader of Statistics Research Section in Bank of Korea, for giving valuable advice and encouragement on this work.

traditional methods rely on surveys to investigate economic sentiment, recent advances in natural language processing (NLP) have made it possible to use unstructured text data to predict economic cycles(Armah et al., 2013; Bennani and Neuenkirch, 2017; Gentzkow et al., 2019; Athey and Imbens, 2019). In this context, this paper proposes a novel Korean news sentiment index (NSI) that measures the economic sentiment of the Korean domestic economy by computing it on a daily basis using news texts scrapped from the internet and a set of NLP techniques. Unlike official economic sentiment indices, which are based on monthly surveys, the proposed index is more frequently and immediately compiled and can help identify inflection points in economic sentiments before the official statistics are released. This paper explains the overall process of building the NSI of Korea using machine learning approaches and provides in-depth assessments of the newly computed index to evaluate its validity and utility from an economics perspective.

News text data is gaining attention as a new source of information due to its rich and diverse content, including vast amounts of economic information on various topics, and its ability to spread information rapidly(Gentzkow et al., 2019; Moon, 2019). Several studies have used news texts in academia and public organizations, including central banks, to extract economic information, such as measuring economic uncertainty(Baker et al., 2016), improving the accuracy of economic cycle forecasting(Bybee et al., 2021; Seki et al., 2022), and predicting inflation expectations(Larsen et al., 2021).

Especially, news sentiment index (NSI) is a widely studied topic in the realm of news text data (Shapiro et al., 2020). NSI is an economic index computed by counting the number of positive and negative sentences in news articles and compiling the difference between the two numbers. Unlike traditional economic sentiment indices, which rely on surveys and are time and cost-intensive, NSI is based on readily available news articles from the internet, providing the advantage of identifying economic sentiment more frequently and immediately with significantly lower costs. However, the process of computing NSI is not straightforward. Unlike survey-based sentiment statistics that undergo strict quality control processes to ensure index stability, extracting economic sentiments from news texts involves various complex problems that must be resolved to use NSI as a reliable economic indicator.

The first challenge of using news text as a source of economic sentiment is that unstructured text data may contain more noise than structured survey data. To mitigate noise in text data, large samples are preferred to compute a stable index. Otherwise, researchers must carefully select journalists who represent public opinion, which is another difficult issue to address. The second hurdle is consistently deciding the economic sentiment of news articles since the interpretation of news articles may differ according to the evaluators. These challenges become more significant when a few evaluators must evaluate large samples of news articles. Hence, it is impractical for humans to classify all news articles to evaluate the economic sentiments of news texts. Conversely, using NLP and text mining technologies makes computing NSI more feasible by providing a way to build a consistent classifier that can efficiently handle large samples of news texts.

The use of NLP and text mining techniques to develop NSI has primarily been investigated by researchers in central banks and international organizations who require quick assessments of the economic situation (Shapiro et al., 2020; Huang et al., 2019; Nguyen et al., 2020; Thorsrud, 2020, 2016; Babii et al., 2021; Jeon et al., 2020; Won et al., 2017; Kim et al., 2021). Central bank researchers have led the development of NSI, aiming to introduce a timely sentiment index as economic sentiments significantly impact the effectiveness of monetary policy (Shapiro et al., 2020). For instance, the Federal Reserve Bank of San Francisco utilizes a lexical approach to analyze news texts and release daily NSI (Shapiro et al., 2020). Similarly, the International Monetary Fund and Reserve Bank of Australia have published works on NSI of different regions based on the lexical approach (Huang et al., 2019; Nguyen et al., 2020). In contrast, the Central Bank of Norway has employed unsupervised learning techniques to determine news indices by topics (Thorsrud, 2020). Academic researchers also actively investigate news texts to improve the accuracy of economic forecasting (Thorsrud, 2016; Babii et al., 2021). For instance, the Bank of Korea has studied text data to develop economic indices using a lexical dictionary for economics, topic modeling, and keyword analysis in news articles (Jeon et al., 2020; Won et al., 2017; Kim et al., 2021, 2019). Although attempts have been made to develop a supplementary sentiment index of Korea using online news data, their work did not employ state-of-the-art

NLP models to analyze Korean texts or conduct in-depth analyses of its correlation with macroeconomic variables and other sentiment indices.

Our proposed NSI differs from other works that use a lexical approach for English text data, as it is computed directly using a machine learning approach. For the rule-based lexical approach, the rules of classifying the sentiments of sentences are expressed by relatively simple patterns that can be explained by words appearing in sentences. The lexical dictionary can also be learned through machine learning as the work in Lee et al. (2019b,a) by learning the sentiments of each word with a statistical model, but by lexical approach, the rule is always expressed by words, which is too restrictive to classify the sentiments of complicated sentences. On the other hand, using the machine learning approach directly on sentences allows us to build more complicated patterns for classification rules and accordingly is likely to increase the prediction accuracy, although the patterns are projected on the new feature space and hence become more difficult to interpret with a simple explanation. Moreover, because there is no known economic sentiment dictionary in Korean such as LM dictionary (Loughran and McDonald, 2011) for English, the machine learning approach can provide an alternative method to build a sentiment classifier.

In order to use a machine learning approach on Korean news text data, we begin by constructing large training samples consisting of randomly chosen news sentences and corresponding labels classified by humans. We then use these training samples to develop a precise classifier for the NSI of Korea. To achieve this, we design a new model based on the transformer neural network classifier, using its encoder structure for classification, and compute the NSI by counting the positive and negative sentiments predicted by the new model.

We assess the validity and utility of the proposed index from multiple perspectives. In terms of validity assessments, we evaluate not only the classification accuracy of the new classifier, but also investigate whether the newly proposed index can reflect the true economic cycle, even prior to the official statistics. For this purpose, we conduct a comparison analysis of the computed monthly NSI with other economic sentiment indices and real economic indices, and examine the impulse response of the sentiment shocks of the

NSI on macroeconomic variables based on a VAR model. We also address the utility of the proposed index. One of the greatest benefits of computing the NSI from news texts is the ability to investigate the reasons why the index fluctuates via its keyword analysis and sector indices. In addition, we investigate the temporal priority of the daily NSI by reviewing cases when the NSI reacts prior to official statistics.

This paper provides contributions in two aspects. First, we provide the NSI of Korea as a regular statistic, accompanied by an in-depth analysis of its validity and utility. The computed daily and monthly NSIs are now publicly available through the Economic Statistics System (ECOS) in Bank of Korea (ecos.bok.or.kr). The NSI has been registered as an experimental statistic in Korea (No. 2022-001) and can be used for future research, such as developing forecasting models using textual data. Second, we present a practical framework for analyzing Korean text data with machine learning models to compile an economic index. The proposed framework is conducted through a fully automated process without human intervention, demonstrating the potential to increase the efficiency of public work for compiling statistics. This framework could also be beneficial for developing new indices in other sectors.

The rest of the paper is organized as follows. In Section 2, we provide a detailed explanation of how we compute the NSI of Korea using machine learning techniques and propose a newly designed sentiment classifier model for Korean texts. In Section 3, we evaluate the validity of the proposed index by comparing it to other economic statistics. In Section 4, we examine the various utilities of the NSI, and finally, in Section 5, we summarize and discuss the future usage of the NSI.

2 Compiling News Sentiment Index (NSI) of Korea

2.1 The Concept of News Sentiment Index

The News Sentiment Index (NSI) is computed by counting the number of positive and negative sentences in daily news articles. Therefore, efficient and accurate classification of sentiment in a vast amount of news articles is crucial for computing a precise NSI. We employ a machine learning approach to achieve this goal. The approach involves analyzing patterns in news sentences that are classified by humans to train a statistical model. The model is then applied to a new sentence to classify its sentiment. The key to this machine learning approach is building a sentiment classifier that fits the data pairs $(s, l)ii = 1^N$ well, where $s_i \in S$ is a news sentence and $l_i = (p_i^{(0)}, p_i^{(1)}, p_i^{(2)})$ is its sentiment label classified by humans, with $p_i^{(0)} = P(s_i \text{ is positive}), p_i^{(1)} = P(s_i \text{ is negative}), p_i^{(2)} = P(s_i \text{ is neutral})$. That is, given a model parameter vector θ, the sentiment classifier f is defined as follows.

$$f_\theta : S \to \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}. \tag{1}$$

2.2 News Text Data

The text data used as input in model (1) consists of news articles collected through web scraping. Web scraping is a method used to directly download publicly available data from the internet. In this study, we collected news articles from an internet news portal that classifies articles into six categories based on authors' choices: politics, economy, society, life/culture, IT/technology, and world. To restrict the scope of the news sentiment we compute to economic sentiment, we collected news articles only from the economy section. We collected all news articles released from 2005 on the news portal. The dataset comprises over 80 publishers and contains 18.3 million articles, with approximately 4,000 articles on normal business days and 1,000 articles on holidays.[1]

Data quality can be a concern when collecting news articles through web scraping as the scraper may collect duplicate articles. To address this issue, we exclude newly collected articles that are exact duplicates of any news in the database within the last 30 days. This helps to ensure that only unique and relevant news articles are included in the database.

[1]

Provider	Num. of articles	Examples
Economic newspapers	10.8m (59.0%)	Money Today, Asisa Economics, EDAILY, etc.
News agencies	4.8m (26.1%)	Yonhan News, Newsis, News1, etc.
Regular news	2.5m (13.8%)	Segye Ilbo, Donga Ilbo, Kyughyang Shinmun, etc.
Others	0.2m (1.1%)	Shindonga, Jungang Sunday, etc.

2.3 Preprocessing for News Text

In order to utilize machine learning on news text data, it's crucial to preprocess the data and convert it into a numerical format that can be interpreted by the sentiment classifier in equation (1). The preprocessing steps include establishing a data structure, tokenizing the text data using part-of-speech (POS) units (Webster and Kit, 1992), and converting the tokenized text data into a numeric format. These processes are achieved through sequential applications of text mining techniques.

The first step in sentiment analysis is to determine the input structure, i.e., whether to classify sentiments based on articles or sentences. Since news articles often contain both positive and negative sentiments, it is not effective to classify the sentiment of an entire article based on its text. To address this issue, the NSI is computed based on randomly sampled sentences from the news database, rather than using entire articles. In other words, the input data consists of each sentence $(s, l)ii = 1^N$ that is randomly selected from the news database.

The second step in making a sentiment classifier understand text data is to tokenize the text in a sentence into POS units. POS is the smallest unit of a word that is assigned according to its syntactic functions in Korean. This process is known as POS tokenization. For instance, the Korean sentence "뉴스심리지수를 작성하였다." is tokenized as follows.

'뉴스심리지수'(proper noun) + '-를'(object case marker) +
'작성'(general noun) + '하'(verb derivative suffix) + '-었'(pre-final ending) +
'-다'(sentence-closing ending) + '.'(period).

Tokenizing sentences into POS makes it possible to distinguish corpus tokens in different conjugated forms by transforming the tokens into the root form. Therefore, POS tokenization is an essential step to use text data for any statistical model. For the collection of sequences of tokens, T, the POS tokenizer is a function, g, assigning a sentence, $s_i \in S$, to a sequence of tokens, $\tau_i \in T$, as follows.

$$g : S \to T. \tag{2}$$

The choice of tokenizer is language-specific and depends on the field in which the tokenizer is used. The tokenizer determines whether a compound word should be split into multiple root words with separate meanings or interpreted as one word. The results of tokenization have a significant impact on the final classification performance. In our work, we use a pre-trained tokenizer called *Open Korea Text (OKT)* (Ryu, 2014) for g, which is available as an open-source software. Our text data for measuring economic sentiments includes various fields and general expressions, and has a large volume. Therefore, we choose the OKT tokenizer, which is widely used to analyze Korean texts for general purposes and shows high efficiency in terms of computational time.

Lastly, the tokenized data is transformed into numerical digits through integer encoding. The integer encoding assigns an integer to each distinct corpus token and represents an input sentence as a numeric vector.To maintain a uniform input data structure, the maximum length of the numeric vector is limited to 80 for each sentence. For sentences shorter than 80 tokens, zeros are padded in front of the numeric vector, and for sentences longer than 80 tokens, the last tokens are truncated. The maximum length, $m = 80$, is chosen based on the 99th percentile of sentence lengths in the news database. The integer encoding assigns a sequence of tokens, $\tau_i \in T$, to an integer vector, $x_i \in \mathbb{R}^m$, through the following bijection function h.

$$h : T \to \mathbb{R}^m \tag{3}$$

The above mentioned preprocessing steps are applied on any sentence input in both training phase for building a sentiment classifier and testing phase for computing daily NSI using the trained model.

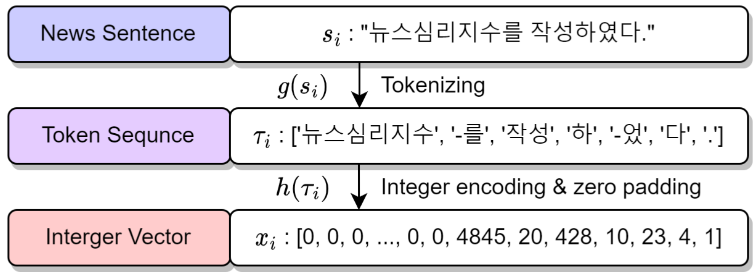


Figure 1: Preprocessing steps of Korean news texts to compute NSI of Korea.

2.4 Construction of Korean News Sentiment (KoNS) Classifier

To build the sentiment classifier (1), we use the state-of-the-art *transformer neural network*, which has been widely adopted in various natural language processing tasks. The transformer model is composed of a multi-head attention and a feed-forward network, and is known to be capable of understanding the context of a sentence (Vaswani et al., 2017) due to its multiple-head attention structure. This structure assigns higher weights to the inputs that require more intensive learning, which enhances the model's ability to capture meaningful relationships among words. We use the encoder structure of the transformer model to build the sentiment classifier for the NSI of Korea, which we call the *Korean News Sentiment (KoNS) classifier*. Figure 2 shows a schematic diagram of the KoNS classifier.

The first step in the KoNS model is to convert the input integer vector x_i of size m into a matrix of real numbers $z_i \in \mathbb{R}^{m \times d_m}$ using the word2vec (Mikolov et al., 2013) mechanism. This mechanism represents each unique word with a d_m-dimensional real-valued vector, so that semantically similar words are located close to each other in the embedded space. The embedded input matrix z_i is then fed into the self-attention layers as the query, key, and values.

The multi-head attention is computed using the equations:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(head_1, \cdots, head_h)\, W^O, && (4)\\ head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), && (5)\\ \text{Attention}(Q', K', V') &= \underset{m}{\text{softmax}}\left(\frac{Q'K'^T}{\sqrt{d_k}}\right) V', && (6) \end{aligned}$$

where d_k is the output dimension of queries, Q', and keys K'; d_v is the output dimension of values, V'; and h is the number of heads. For the output dimension of the embedding, d_m, $W_i^Q \in \mathbb{R}^{d_m \times d_k}$, $W_i^K \in \mathbb{R}^{d_m \times d_k}$, $W_i^V \in \mathbb{R}^{d_m \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_m}$ are the model parameters. Since KoNS is a classification model, we used self-attention mechanism, i.e., the embedded input sequence, z_i, is used for all Q, K, and V. We set $d_k = d_v = d_m = 32$ and $h = 2$. Furthermore, Concat is a function that concatenates the input matrices along the second axis of the dimensions such tat $\text{Concat}(x_1, \cdots, x_h) = [x_1^T \cdots x_h^T]^T$. $\underset{m}{\text{softmax}}$ is a real-valued function that normalizes the input vector such that the output is a probability distribution over the elements such that $\underset{m}{\text{softmax}}(x_j) = \frac{e^{x_j}}{\sum_{l=1}^{m} e^{x_l}}$, $\underset{m}{\text{softmax}} : \mathbb{R} \to (0, 1)$ for $j = 1, \cdots, m$ where the m is the length of the input sequence.

The output of the multi-head attention is attached to the original embedded input, and followed by layer-normalization (Ba et al., 2016) to reduce computation time. The output is then fed into a feed-forward network with its output dimension equal to $d_f = 32$ in the transformer block. The equation for the feed-forward network is:

$$\text{FeedForward}(x) = \sigma(W^F x + b^F), \tag{7}$$

Here, $W^F \in \mathbb{R}^{d_m \times d_f}, b^F \in \mathbb{R}^{d_f}$, and $\sigma(\cdot)$ is the rectified linear unit (ReLU) activation function. The output of the previous layer is denoted by x in (7).

The output of the transformer block is then averaged along the first axis, which corresponds to the m number of words, and passed through two fully-connected dense layers. The output dimensions of these layers are set to 20 and 3 to map the output to the three classes: positive, negative, and neutral. Finally, a softmax function is applied over the 3 classes to obtain the probability distribution for each class.

The configuration of the transformer neural network, including its hidden unit dimensions (d_k, d_v, d_m, d_f), is important for the performance of the sentiment classifier. In our work, we follow the conventions used for the transformer model for similar text classification tasks. While it is known that a large number of hidden layers and units are required to express a complex function (LeCun et al., 2015), there is no theoretical reason to use more than two layers (Heaton, 2008). Therefore, we choose to use a comparable configuration for similar tasks and data(Chollet et al., 2015). For more detailed aspects of the transformer neural network used to build KoNS, we refer the reader to the original paper by Vaswani et al. (2017).

Now, we express KoNS in (1) as the model $\hat{f}_\theta$ taking a news sentence $s_i \in S$ as input and predicting the probability of the sentiment, $\hat{l}_i$, via the sequential procedures of preprocessing and conducting the sentiment classification. That is,

$$\begin{aligned} \hat{l}_i &= \hat{f}_\theta(s_i) && (8) \\ &= \text{TransformerBasedClassifier} \circ h \circ g(s_i) && (9) \end{aligned}$$

where $\hat{l}_i$ is the predicted value such that $\hat{l}_i \in \{(\hat{p}_i^{(0)}, \hat{p}_i^{(1)}, \hat{p}_i^{(2)})\}$, and $0 \leq \hat{p}_i^{(c)} \leq 1$ for $c = 0, 1, 2$, and $\sum_{c=0}^{2} \hat{p}_i^{(c)} = 1$, and h and g indicate the integer embedding process and tokenizer respectively.

The model parameters, θ, are estimated by minimizing the cross-entropy loss function so that the difference between the observed label, $\{(p_i^{(0)}, p_i^{(1)}, p_i^{(2)})\}$, and the predicted probability, $\{(\hat{p}_i^{(0)}, \hat{p}_i^{(1)}, \hat{p}_i^{(2)})\}$, becomes small for each sentence, $s_i, i = 1, \cdots, N$.

$$L(\theta | \{(s_i, l_i)\}_{i=1}^{N}) = -\sum_{i=1}^{N} \sum_{c=0}^{2} p_i^{(c)} \log \hat{p}_i^{(c)}, \tag{10}$$

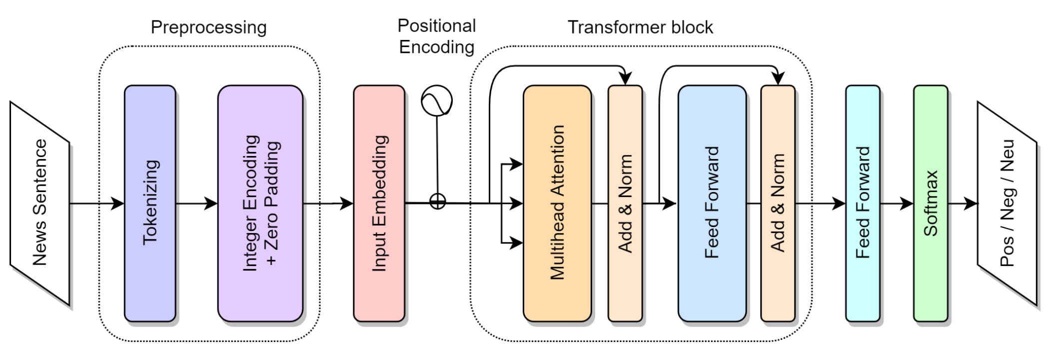


Figure 2: A schematic diagram of the transformer-based sentiment classifier for NSI of Korea.

In addition to the KoNS sentiment classifier, we build another model with the same structure of (1), but this time, to classify the geographical scope of the sentences into one of the three categories based on their contents: domestic, foreign, and both combined. We use only the sentences classified as domestic to compute NSI of Korea. The purpose of this process is to exclude foreign news articles from the compilation of NSI of Korea as they often show different patterns from the domestic economic sentiment. For simplicity of explanation, we assume hereafter that any data inputs for KoNS are cleaned with the geographical scope classifier in advance and consist of only domestic sentences.

2.5 Training Data and Model Training

To estimate KoNS, it is crucial to have a training dataset with correct labels. We combined data from three separate projects to create a training dataset, which comprised a total of $446,478$ sentences. In the first project, conducted in 2019, we labeled $230,583$ news sentences sampled from publications between January 2008 and December 2018. In 2020, we added $84,069$ more sentences to the dataset based on news published between January 2008 and May 2020. In our most recent effort, we generated an additional $131,826$ labeled sentences using news published between January 2005 and June 2021. We ensured that there were no duplicate sentences in the training data.

The training sentences were subjected to sentiment classification by 16 trained individuals who were hired explicitly for the labeling task. Each sentence was classified into one of three categories, namely positive, negative, or neutral sentiment. To minimize measurement errors, we enlisted additional reviewers, economists from the Bank of Korea, to review the sentiment labels after the initial classification.

However, given the complexities of economic subjects and contexts, certain sentences presented challenges in their sentiment classification. For example, inflation can have both positive and negative impacts on the economy, rendering it difficult to classify as either positive or negative. In response to this issue, we established minimal guidelines for frequently occurring economic subjects, which we outline in Table 1. These guidelines were designed to provide a framework for the sentiment classification of economic subjects that present challenges, thereby promoting consistency in the labeling process.

Subjects	Guidelines
Stock price, interest rates, exchange rates, and other asset prices	• A sentence simply reading the change of the index is neutral. • If the change is mentioned with specific reasons, the sentence is classified as positive or negative depending on its context.
Industry and companies	• Advertisements are neutral. • A simple statement on historical facts is neutral. • If past events are mentioned with precise implications for the present situation or the future expectations, the sentence is classified as positive or negative. • Despite a brief remark, if a sentence mentions the change in financial statements, it is classified as positive or negative.
Real estate and construction	• Advertisements for apartment sales are neutral. • Statements on the unsold apartments are classified as positive or negative.
Government and public sectors	• Government supports are classified as positive or negative. • Government regulations are classified as positive or negative when specific expected impacts are indicated.

Table 1: Minimal guidelines for sentiment classification of economic news sentences to build the training data. The subjectivity in sentiment interpretation is unavoidable because the economic sentiment is indeed abstract.

Table 1 provides guidelines that limit the interpretation of sentiment and result in a more cautious classification by highlighting the clear tones found in the training data. Consequently, around 80% of the sentences in the training data are classified as neutral, while the remaining 20% are divided equally between positive and negative classifications. Economic sentiment is an abstract concept, and therefore, subjectivity in sentiment interpretation is inevitable when constructing training data for machine learning approaches. Even with traditional lexical approach and surveying methods, subjectivity is present as individuals use different criteria to evaluate their economic situation.

It is worth noting that the imbalanced class weights of the training data can affect the final prediction of a classifier. However, since we do not possess the labels of the daily news data, we cannot guarantee that the class weight of the daily news is consistent with the training dataset. To address this issue and account for the fluctuation of the computed NSI, we modify the loss function in (10) by incorporating class weights, $w^{(c)} = \frac{N}{3N^{(c)}}$, where $N^{(c)}$ denotes the number of sentences in class c of the training data.

$$L(\theta|\{(s_i, l_i)\}_{i=1}^N) = -\sum_{i=1}^{N} \sum_{c=0}^{2} w^{(c)} p_i^{(c)} \log \hat{p}_i^{(c)}, \tag{11}$$

The comparison between using the loss with and without the weight adjustment is demonstrated in Table 2 in Section 3.

2.6 Prediction for Daily News Data

Finally, to compile the NSI, we apply the KoNS classifier to daily news data. We collect around 4,000 news articles from the internet per weekday as of 2021, which is equivalent to about 70,000 sentences. The number of sentences decreases to around 10,000 per day on weekends. To reduce computing costs and stabilize the NSI across days, we randomly sample 10,000 sentences every day and use only the sampled sentences to compute the NSI. This sampling process prevents certain days from dominating the economic sentiment of the NSI when more news articles are released on those days.

2.7 Computing News Sentiment Index (NSI) of Korea

The daily NSI is calculated by first predicting the sentiment of daily news sentences using KoNS. We then count the number of positive and negative sentences from the sample collected during the previous seven days. The daily NSI is computed as the ratio of the difference and sum of the positive and negative counts. Using a seven-day period of data ensures a smooth and stable index. Similarly, the monthly NSI is computed by the same procedure but based on news sentences collected during the corresponding month.

To facilitate comparison with past economic sentiment, both daily and monthly NSIs are standardized using their long-term averages. The NSIs are standardized such that their averages and standard deviations are both equal to 100 and 10, respectively. The standardization interval starts from 2005 and extends to the end of the previous year at the beginning of each new year. The computation process for the CCSI and ESI is also standardized in the same manner.

Specifically, the daily NSI of a particular day t, denoted as $NSI_t^{(\text{daily})}$, is calculated using the following formula.

$$NSI_t^{(\text{daily})} = \left(\frac{X_t - \bar{X}}{S}\right) \times 10 + 100, \tag{12}$$

$$\text{where} \quad X_t = \frac{\sum_{u=1}^{7} P_{t-u} - \sum_{u=1}^{7} N_{t-u}}{\sum_{u=1}^{7} P_{t-u} + \sum_{u=1}^{7} N_{t-u}}, \tag{13}$$

$$\bar{X} = \frac{1}{|U|}\sum_{u\in U} X_u, \quad S = \sqrt{\frac{1}{|U|-1}\sum_{u\in U}(X_u - \bar{X})^2}. \tag{14}$$

Here, U represents the index set of days for the standardized interval, which is corresponding to the days between January 1st, 2005 and the last day of the previous year. Without loss of generality, the monthly NSI, $NSI_v^{(\text{monthly})}$, for a particular month, v, is compiled in the same way, but by replacing U with the index set of months V which includes all months between January 2005 and December of the previous year, and X_t with X_v as follows.

$$X_v = \frac{\sum\limits_{v \in M_v} P_v - \sum\limits_{v \in M_v} N_v}{\sum\limits_{v \in M_v} P_v + \sum\limits_{v \in M_v} N_v}, \tag{15}$$

where M_v is the index set of days belonging to the particular month v. That is, the monthly NSI is calculated by analyzing the sentences published during the corresponding month, whereas the daily NSI is calculated using news data collected during the previous seven days. This is because the daily NSI can be highly volatile, and using a seven-day period of data helps to generate a more stable index.

A NSI greater than 100 indicates that the economic sentiment in the analyzed news articles is more optimistic than the past average, while a NSI less than 100 indicates a more pessimistic sentiment.

2.8 Automated Compilation

NSI has been designed to be compiled automatically every week without the need for human labor. The entire process, from web-scraping of daily news articles from the internet to text preprocessing, sentiment prediction, and compilation of the daily and monthly NSIs, is carried out using a Python script that runs on an automated batch. The script collects news articles from the internet every day, from the last point of the database to the previous day at 5 AM, and then uses KoNS to compute the daily and monthly NSIs. This automation of the compiling process improves efficiency, saves labor, and reduces costs compared to the survey approach.

3 Validity Assessments of NSI

In this section, we evaluate the newly computed NSI of Korea from multiple perspectives. Clearly the validation of NSI is not only focused on the accuracy of the sentiment classification, but also on whether it implies the true economic sentiment and even indicates the truth prior to other official statistics. Therefore, we validate NSI by sentiment classification accuracy, comparative analysis with other economic indicators based on the cross-correlation, Granger causality tests, and impulse response analysis using a VAR macroeconomic model.

3.1 Sentiment Classification Accuracy

We compared the accuracy of the KoNS classifier against the lexical approach and other statistical models using 10-fold cross-validation[2]. To implement the lexical approach, we translated the sentiment of 50,754 English keywords from Shapiro et al. (2020) into their corresponding Korean words, as there is no sentiment dictionary for economic keywords in Korean. Table 2 shows the classification accuracy of KoNS and other standard methods. We compare only the positive and negative classes because the neutral class does not affect the computation of NSI. Table 2 indicates that KoNS outperforms the other methods that were evaluated. The lexical approach demonstrates inferior accuracy compared to the machine learning approach, even when using simple linear regression. One interesting finding is that the simple linear regression achieves a competitive accuracy to the more sophisticated NLP models. This appears mainly because the news text sentences are relatively straightforward and their sentiment is easily revealed based on the specific words in sentences rather than being influenced by complicated sentence structures or contextual implications.

[2] The performance metrics have been defined as follows. Denote the true sentiment label and predicted sentiment label of ith sentence by l_i and $\hat{l}_i$ respectively. Since we are concentrating on the case where the sentiment of the sentence is either positive or negative, $\hat{l}_i, l_i$ take values from the set $\{\text{positive}, \text{negative}\}$. Our performance metrics include accuracy, sensitivity, specificity, and precision, which we define as follows:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}} \\
\text{Sensitivity} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \\
\text{Specificity} &= \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \\
\text{Precision} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\end{aligned}$$

where the term 'True Positives' refers to the number of samples where both the predicted label ($\hat{l}_i$) and the true label (l_i) are positive. Conversely, 'False Positives' denotes the number of samples where the predicted label is positive ($\hat{l}_i$), but the true label is negative (l_i). Similarly, the terms 'True Negatives' and 'False Negatives' are defined based on the number of samples where both the predicted label and the true label are negative or where the predicted label is negative, but the true label is positive, respectively. The F1 score is a balanced performance metric of classification which defined as a harmonic mean of sensitivity and precision that is $\left[\frac{1}{2}\left(\frac{1}{\text{Sensitivity}} + \frac{1}{\text{Precision}}\right)\right]^{-1}$.

[3] Feed-forward network (FFN) is a specific structure of neural networks that consist of layers connected by an affine transformation and an activation function without any recurrent connection.

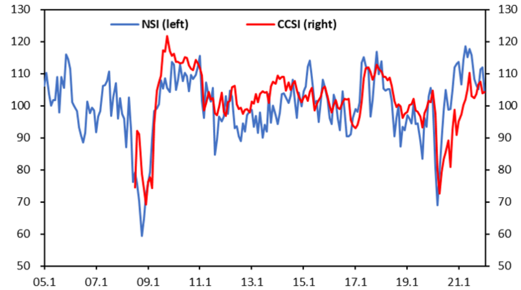
	LEX	LR	FFN	SVM		KoNS	
		SL	SL	SL	WAL	SL	WAL
Accuracy	0.82	0.85	0.85	0.86	0.81	0.97	0.92
	(0.027)	(0.016)	(0.015)	(0.026)	(0.017)	(0.009)	(0.022)
Sensitivity	0.81	0.84	0.90	0.81	0.91	0.96	0.91
	(0.055)	(0.043)	(0.042)	(0.038)	(0.051)	(0.019)	(0.035)
Specificity	0.83	0.86	0.76	0.86	0.80	0.98	0.93
	(0.024)	(0.032)	(0.061)	(0.044)	(0.032)	(0.008)	(0.014)
Precision	0.78	0.85	0.83	0.85	0.80	0.98	0.93
	(0.067)	(0.032)	(0.034)	(0.044)	(0.043)	(0.012)	(0.029)
F-1 score	0.80	0.85	0.79	0.86	0.80	0.98	0.93
	(0.036)	(0.020)	(0.026)	(0.037)	(0.021)	(0.009)	(0.020)

Table 2: 10-fold cross-validation accuracy for the binary classification of the positive and negative classes, i.e., the neutral class has been ignored in both predicted and true labels. LEX: lexical approach, LR: logistic regression, FFN: feed-forward network[3], SVM: support vector machine, SL: standard loss used, WAL: weight-adjusted loss used.

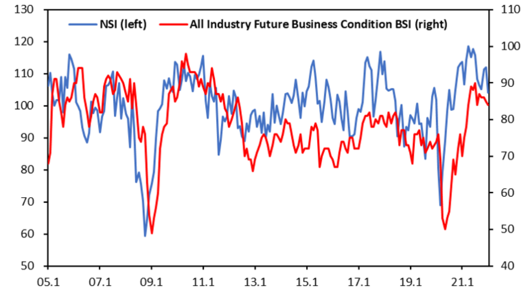
3.2 Comparative Analysis with Economic Indicators

The computed NSI is compared with various official statistics. Firstly, the following economic sentiment indicators are compared to NSI: consumer survey index (CSI), composite consumer sentiment index (CCSI), business survey index (BSI) representing the economic sentiment of entrepreneurs, and economic sentiment index (ESI) which is a composite index of CSI and BSI. Because these economic sentiment indicators are compiled based on monthly surveys, we compare them to monthly NSI. The comparative analysis is conducted based on the available data until December 2021. Note that BSI and ESI are available since January 2005, and CCSI and CSI since July 2008. In Table 3, it is demonstrated that the monthly NSI is leading most of the economic sentiment indicators by 1 to 2 months with high correlation according to the cross-correlation analysis. Especially, the monthly NSI shows the high correlation of 0.75 with CCSI leading it by 1 month. NSI is also compared to real economic indicators. The monthly NSI is leading the cyclical composite leading index (CCLI) by 2 months with its correlation equal to 0.76. The quarterly NSI is coincident with the gross domestic product (GDP), which is measured by quarter-on-quarter (QoQ) change of the seasonally-adjusted real GDP, with its correlation equal to 0.53.

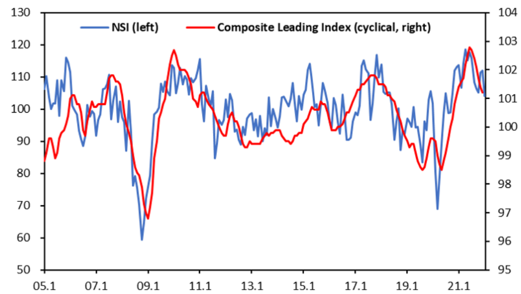
Figure 3 shows that the monthly NSI has the similar trend to CCSI and all industry



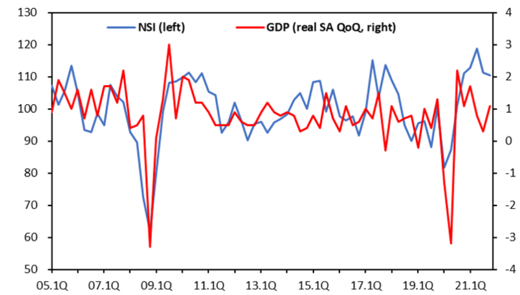
(a) Monthly NSI and CCSI



(b) Monthly NSI and BSI



(c) Monthly NSI and CCLI



(d) Quarterly NSI and GDP

Figure 3: Comparisons between NSI and various economic indicators.

Economic indicators		Max. Corr. (Lag) KoNS	Max. Corr. (Lag) LEX
CCSI		0.75 (-1)	0.62 (-1)
ESI		0.61 (-2)	0.27 (-1)
CSI	Living Standard of Household	0.74 (-1)	0.80 (-1)
	Domestic Economic Situation	0.73 (-1)	0.54 (-1)
	Expectation of Living Standard	0.73 (-1)	0.63 (-1)
	Expectation of Domestic Economic Situation	0.70 (-1)	0.45 (0)
	Expectation of Household Income	0.68 (-1)	0.64 (-1)
	Spending Plan	0.57 (-1)	0.42 (-1)
	Expectation of Employment Situation	0.72 (-1)	0.54 (-1)
BSI	All Industries Business Condition	0.64 (-1)	0.23 (-1)
	All Industries Profitability	0.68 (-1)	0.49 (-1)
	All Industries Financial Situation	0.64 (-1)	0.33 (-1)
	All Industries Future Business Condition	0.61 (-2)	0.20 (-2)
	All Industries Future Profitability	0.65 (-2)	0.44 (-2)
	All Industries Future Financial Situation	0.61 (-2)	0.29 (-2)
Real Indices	KOSPI(monthly closing price YoY%)	0.68 (-1)	0.26 (-1)
	Cycle of Composite Leading Index(CCLI)	0.76 (-2)	0.36 (-1)
	GDP (real SA QoQ%)	0.53 (0)	0.32 (0)

Table 3: table
Cross correlation analysis results between the NSI and major economic indicators.

future business condition BSI. In addition, we can see in the figure that the NSI hits the lowest point 1 to 2 months earlier than the official statistics during the global financial crisis in 2008 and the COVID-19 crisis in 2020.

3.3 Granger Causality Tests

Furthermore, we computed NSI using the LEX approach and compared it to NSI calculated using the KoNS classifier. As shown in Table 3, the correlations of NSI to other sentiment indicators and macroeconomic variables were significantly low for the lexical approach. While the lexical approach exhibited a relatively high correlation with the Composite Consumer Sentiment Index (CCSI), it displayed weak correlations with the Business Survey Index (BSI) and real indices. This limitation is due to the lexical approach's use of a restricted set of keywords to describe business conditions.

To test if the NSI leads economic indicators, we also performed the Granger causality test on those economic indicators that appear to be led by NSI from the cross-correlation analysis. For fair comparison, we only consider data between July 2008 and December 2021, as CCSI and CSI data are not available before July 2008. To determine whether variable x Granger-cause variable y, we use the following model:

$$y_t = \alpha_0 + \sum_{l=1}^{L} \alpha_l y_{t-l} + \sum_{l=1}^{L} \beta_l x_{t-l} + \varepsilon_t.$$

The model is estimated using OLS and causality is tested under the null hypothesis $\beta_l = 0$ for all l. The decision rule is if the null hypothesis is rejected or not. That is, if the null hypothesis is not rejected, it means x does not seem to cause y. The test is performed with the Wald test procedure and L is set to 2.

Table 4 summarises the test results. It provides a statistically significant evidence for the Granger causality of the NSI to economic indicators, except for few (KOSPI and CCLI). For most economic indicators, especially for the economic sentiment related indicators, the null hypothesis that "the NSI does not cause the economic indicator" is rejected at the 0.1% significance level, while the null hypothesis that "the NSI does not cause the economic indicator" it is not rejected at the 1% significance level. This result is robust to alternative specifications, where L is set to 1 or 3.

However, it is important to note that the Granger causality test only measures the statistical predictability of one variable over another, and does not imply that news articles are the direct cause of variations in economic sentiment. Identifying the channels through which news influences the beliefs of economic agents is beyond the scope of this paper. Nevertheless, the results support the view that the NSI is useful for predicting economic sentiment.

Y	N of obs.	Null hypothesis	
		NSI $\nRightarrow$ Y	Y $\nRightarrow$ NSI
CCSI	160	45.5*** (< 0.1%)	1.6 (0.46)
ESI	160	36.8*** (< 0.1%)	1.4 (0.49)
Living Standard of Household	160	45.3*** (< 0.1%)	0.6 (0.75)
Domestic Economic Situation	160	34.2*** (< 0.1%)	0 (0.98)
Expectations of Living Standard	160	35.1*** (< 0.1%)	1 (0.59)
Expectations of Domestic Economic Situation	160	23.8*** (< 0.1%)	2.4 (0.3)
Expectations of Household Income	160	48.1*** (< 0.1%)	6.8** (0.03)
Spending Plan	160	36.2*** (< 0.1%)	4.1 (0.13)
Expectations of Employment Situation	160	26.6*** (< 0.1%)	0.9 (0.65)
All Industuries Business Condition	160	9.1** (0.011)	2.5 (0.28)
All Industuries Profitability	160	18.2*** (< 0.1%)	1.2 (0.55)
All Industuries Financial Situation	160	15.3*** (< 0.1%)	0.9 (0.63)
All Industuries Future Business Condition	160	46.1*** (< 0.1%)	1.3 (0.53)
All Industuries Future Profitability	160	41.2*** (< 0.1%)	1.9 (0.39)
All Industuries Future Financial Situation	160	53.6*** (< 0.1%)	0.3 (0.84)
KOSPI(monthly closing price YoY%)	160	3.1 (0.22)	3.8 (0.15)
Cycle of Composite Leading Index (CCLI)	160	16.3*** (< 0.1%)	10.2*** (0.01)

Table 4: Summary of Granger Causality Test Results(TBD)

Note: The elements of third and forth columns are the value of χ^2 statistic under given null hypothesis. Here, *, **, and *** are 10%, 5%, and 1% significance level respectively, and the values in parenthesis are the corresponding p-values.

3.4 Impulse Response Analysis

We validate the impact of economic sentiment shocks measured by NSI using a macroeconomic model and comparing it to ESI shocks. We build a standard VAR system following the procedure of van Aarle and Kappler (2012), in which they study the interaction between confidence indicators and macroeconomic adjustments using four variables: unemployment (UNE), industrial production (IND), retail sales (RET), and economic sentiment index (ESI). In our study, we use monthly unemployment rate, industrial production index, and retail business service index for the variables. All the variables are seasonally differenced by using year-on-year (YoY) growth rates. Let t denote time. Then, the structural representation of the considered model is as follows.

$$Cy_t = \alpha + \sum_{i=1}^{k} C_i y_{t-1} + \epsilon_t, \tag{16}$$

$$\text{where} \quad y_t = \begin{pmatrix} IND_t \\ UNE_t \\ RET_t \\ NSI_t \end{pmatrix}, \quad \epsilon_t = \begin{pmatrix} \epsilon_t^{IND} \\ \epsilon_t^{UNE} \\ \epsilon_t^{RET} \\ \epsilon_t^{NSI} \end{pmatrix}.$$

Here, y_t denotes the vector of endogenous variables, and ϵ_t denotes the vector of residuals. We assume C^{-1} has the recursive structure and its reduced form errors $e_t = C^{-1}\epsilon_t$, which indicates C^{-1} has the lower trianglular structure.

	ESI model			NSI model		
	IND	UNE	RET	IND	UNE	RET
SI_{t-1}	0.314*	0.075*	0.082**	0.012**	-0.018*	0.046**
	(0.060)	(0.089)	(0.035)	(0.044)	(0.063)	(0.025)
SI_{t-2}	-0.181*	-0.114*	-0.043**	0.115**	0.056*	-0.004**
	(0.060)	(0.090)	(0.036)	(0.044)	(0.064)	(0.025)
Adj. R^2	0.586	0.526	0.321	0.557	0.524	0.324

Table 5: Estimation results of VAR model for macroeconomic variables with sentiment indicators(SI): NSI and ESI. *, ** indicate the statistics are significant with $\alpha = 0.10$ and 0.05 respectively.

We recover the orthogonalized shocks with $\epsilon = Ce_t$ where $C = chol(\Sigma)$, which is the Choleski decomposition of the covariance matrix of residuals. We checked the robustness of the model by changing the ordering of the variables but the estimated impulse responses have little change. Table 5 shows the estimation results of the models with NSI and ESI. The table demonstrates that NSI has very similar explanatory power to ESI and has the adjusted

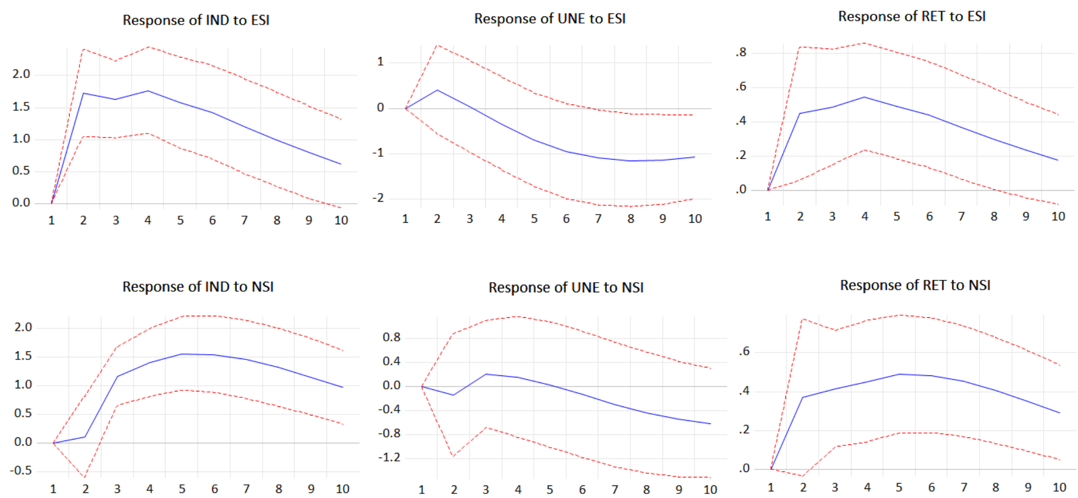


Figure 4: Impulse responses of macroeconomic variables on the economic sentiment shocks measured by ESI and NSI.

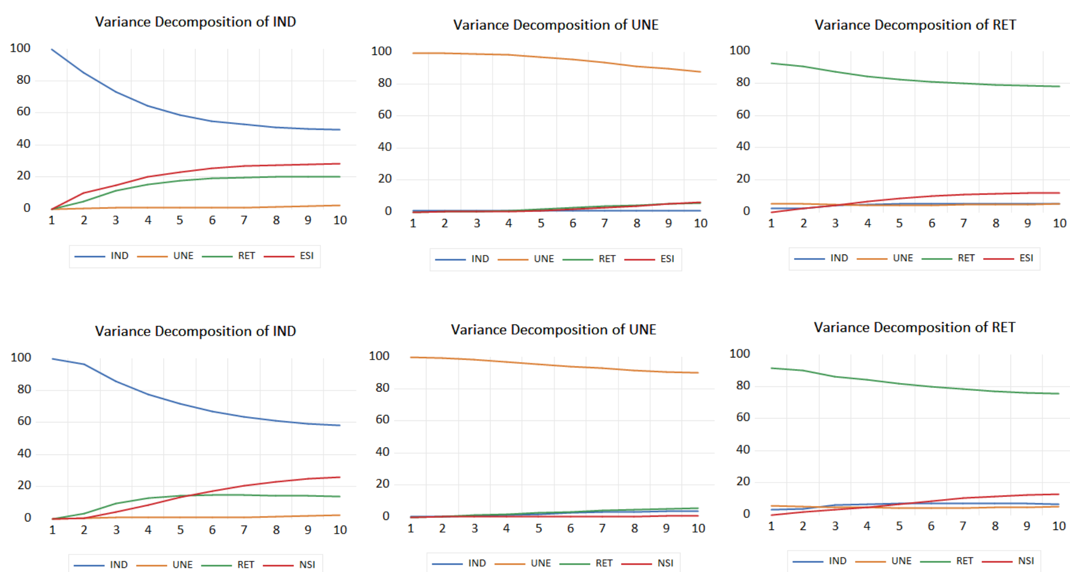


Figure 5: Variance error decomposition for the VAR models with ESI and NSI. The sentiment shocks measured by ESI or NSI largely contribute to the variance in the business cycle variable, IND.

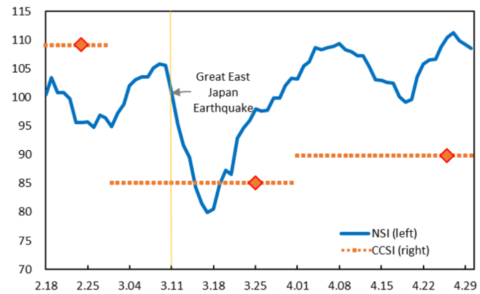
R^2 close to that of ESI model. The impulse responses of the two models are displayed in Figure 4. The impulse response of the industrial production (IND) to NSI implies that the production is boosted by the increase of economic confidence with its hike appearing in 3 months. Whereas the ESI shocks on the industrial production shows the largest hike in 2 months, which supports the same result that NSI is leading ESI. The variance error decompositions, displayed in Figure 5, also indicate that the sentiment shocks measured by NSI largely contribute to the variance in the business cycle variable, and for the other two economic variables, the impact of NSI is less clear, which coincides with the findings in van Aarle and Kappler (2012) who come to the similar conclusion using ESI in the euro area.

4 Utility Assessments of NSI

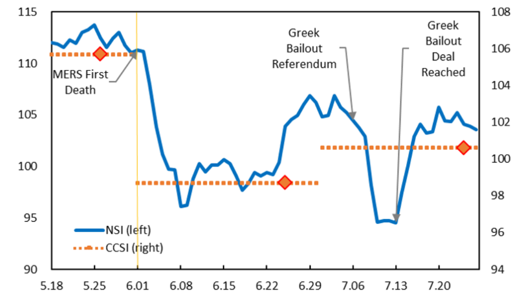
In this section, we present utility assessments of NSI. One of the most significant advantages of NSI is its timeliness as a daily economic indicator. To test the temporal priority of daily NSI, we compare its ability to identify inflection points in economic sentiment to that of other official statistics. Additionally, NSI provides greater explainability by providing detailed information about probable factors of fluctuation through keywords and metadata. We analyze NSI using keyword analysis and sectoral NSIs computed for three economic sectors.

4.1 Temporal Priority

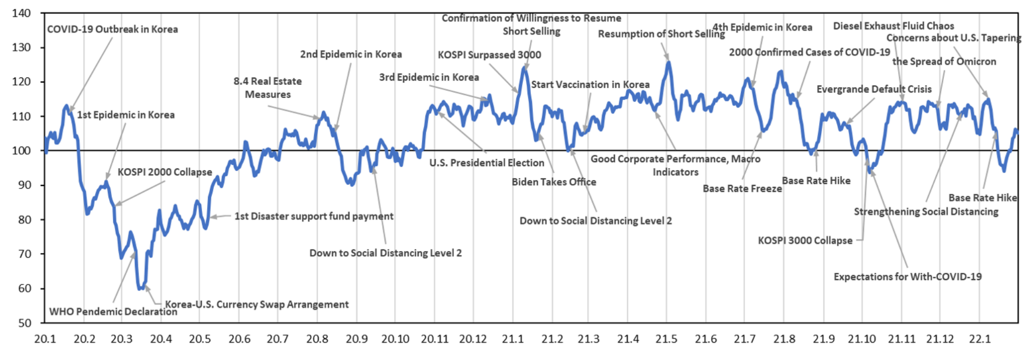
NSI has the advantage of being compiled on a daily basis, allowing it to quickly identify changes in economic sentiment before monthly official statistics based on surveys are released. As shown in Figure 6, NSI can immediately quantify the impact of important economic events and effectively detect inflection points in economic sentiment. For example, in 2011, when the Great East Japan Earthquake struck in early March, NSI dropped sharply and then rebounded soon after the event, while the drop in the Consumer Confidence Survey Index (CCSI) was observed at the end of March when the statistics were released after the monthly survey was conducted. Similarly, in 2015, NSI dropped immediately after the first death from MERS was known to the public on June 1st, but CCSI could not indicate the event until the end of June. NSI thus provides immediate information earlier than survey-based official statistics and can be used as a supplementary indicator to quickly detect inflection points in economic sentiment.



(a) Great East Japan Earthquake struck in March 11, 2011



(b) MERS first death occurred in June 1, 2015



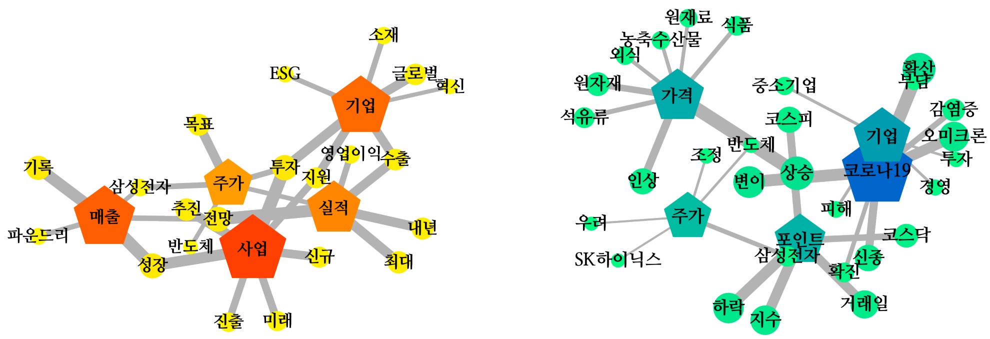
(c) NSI around COVID-19 pandemic in 2020 and 2021

Figure 6: NSI with major economic events. In (a,b), the red bullets indicate the day monthly CCSI is released. These figures demonstrate that NSI detects the inflection points of economy effectively and even prior to official monthly statistics.

4.2 Explainability and Keyword Analysis

Another advantage of NSI is that the analysis of keywords reveals the factors behind NSI fluctuations. Figure 7 displays the keyword networks in positive and negative news articles, which were classified by KoNS in the last week of November 2021. The positive keywords of the period mainly consisted of company earnings, while the negative keywords were related to the price hike of raw materials and the COVID-19 Omicron mutation. The keywords can also disclose the main economic issues at a given time. For instance, in the last week of

November 2021, ‘Omicron’ mutation, ‘KOSPI,’ and the government’s deliberation on the ‘budget’ for supporting small business owners were mentioned more than in the previous week, indicating the main economic issues of the week. Table 6 presents the keywords of the same week in positive and negative news articles divided by sub-sectors: macro, finance, and industry. The table illustrates that the impact of stock market fluctuations was mentioned in both positive and negative articles in the finance sector. In the macro sector, the rise of exports in November was largely responsible for the rise of NSI, while the worsening industrial production statistics were attributed as a factor for the decline in NSI. Such detailed information on economic sentiments is challenging to obtain through surveys.



(a) Positive keywords (b) Negative keywords

Figure 7: The keyword networks of positive and negative sentences classified by KoNS. The pentagon nodes indicate the 5 top main keywords appearing in the sentences in the last week of November 2021, and the circle nodes indicate the related keywords appearing together with the main keywords.

4.3 Sector NSIs

NSI has the advantage of computing sector-wise sentiment easily by dividing news sources into sectors. This is a benefit of using news articles to compute economic sentiment since sector information is obtained as metadata when collecting the news articles. Internet portals categorize news articles, and hence, we can use these categories to divide news sources into three sub-categories: macro, finance, and industry, and compute sector NSIs. As shown in Figure 8, although the sector NSIs are highly correlated with the aggregated NSI, they exhibit different patterns at particular time points when specific events occur, such as

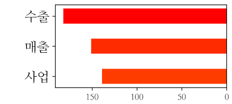
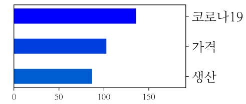
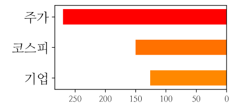
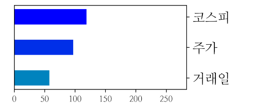
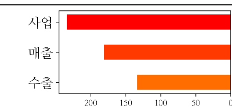
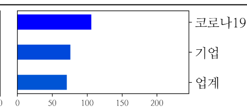
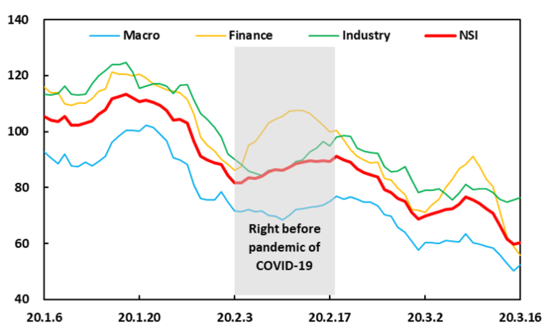
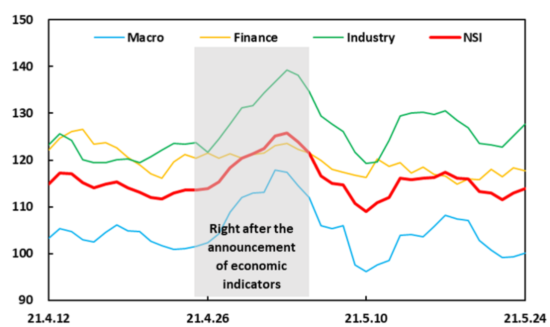
Sec.	Positive sentences		Negative sentences	
	Related words	Keywords	Keywords	Related words
Mac.	(증가, 11월, 최고, 실적) (분기, 증가, 기록, 성장) (투자, 지원, 글로벌, 성장)			(확산, 변이, 오미크론, 위기) (상승, 석유류, 외식, 원자재) (감소, 자동차, 광공업, 반도체)
Fin.	(상승, 전망, 실적, 삼성전자) (지수, 외국인, 상승, 코스닥) (성장, 투자, 확대, 미래)			(하락, 지수, 오미크론, 최저) (하락, 삼성전자, 업황, 우려) (포인트, 하락, 코스피, 지수)
Ind.	(투자, 추진, 소재, 친환경) (분기, 증가, 성장, 파운드리) (증가, 반도체, 11월, 실적)			(확산, 신종, 장기, 변이) (피해, 중소기업, 과징금) (항공, 반도체, 조선, 자동차)

Table 6: The most appeared keywords and their related words in news sentences grouped into one of the positive and negative sentiments in three sectors: macro, finance, and industry. The bar plots indicate the counts of keywords appearing in the groups of sentences. The data is as of the last week of November 2021.

the COVID-19 outbreak and company earnings announcements. Sector NSIs provide more detailed information on economic sentiment in different sectors. Table 7 shows that the sector NSIs became more homogeneous after the COVID-19 outbreak, suggesting that the COVID-19 pandemic dominated the factors in economic sentiment fluctuations after its outbreak. Obtaining such information through surveys would significantly increase costs by adding more questions.



(a) COVID-19 outbreak in Mar. 2020



(b) 1Q earning season in Apr. 2021

Figure 8: Sector NSIs with economic events. The sector NSIs provide more detailed information about the economic sentiments. Right before the COVID-19 pandemic in March 2020, finance sector NSI increased high due to the earning surprises of financial holdings companies in Korea; Nevertheless other sector NSIs remained flat due to the concern for the new virus. In April 2021, macro and industry sector NSIs rose after 1Q GDP of Korea was announced higher than anticipated.

Sector	After 2015			After 2018			After 2020		
	Mac.	Fin.	Ind.	Mac.	Fin.	Ind.	Mac.	Fin.	Ind.
Macro	1.00	0.75	0.76	1.00	0.85	0.84	1.00	0.88	0.92
Finance	-	1.00	0.67	-	1.00	0.71	-	1.00	0.79
Industry	-	-	1.00	-	-	1.00	-	-	1.00
Aggregated	0.94	0.89	0.88	0.97	0.91	0.91	0.98	0.92	0.95

Table 7: The correlation coefficients between the sector NSIs and the aggregated NSI. The correlations are computed based on the daily indices.

5 Summary and Discussion

In this paper, we present the development and evaluation of a news sentiment index (NSI) for Korea, which is computed based on news articles collected from the internet using web-scraping techniques. To compute NSI, we develop a Korean news sentiment (KoNS) classifier using a state-of-the-art natural language processing (NLP) model. The encoder structure of the transformer model is utilized, and the model is trained with a modified loss function to control imbalanced data. Over 450,000 training sentences labeled by 16 trained personnel are used to train KoNS. Daily and monthly NSIs are computed based on the counts of the positive and negative sentences classified by KoNS for daily news sentence samples.

The validity of the computed NSI is evaluated in multiple perspectives. The cross-correlation analysis shows that NSI leads the composite consumer sentiment index (CCSI) by one month with a correlation of 0.75 and the cycle of the composite leading index (CCLI) by two months with a correlation of 0.76. Additionally, the impulse response analysis shows that a rise in NSI stimulates industrial production, with the hike appearing in three months.

Furthermore, we evaluate the various utilities of NSI. Daily NSI is found to correctly detect inflection points in economic sentiments before official statistics based on monthly surveys. Keyword analysis reveals the factors that cause NSI fluctuations and provides more information than can be quantified through survey-based statistics. Lastly, sector NSIs are easily computed by dividing news articles into sectors, and they provide more detailed information on economic sentiments in different sectors.

Our contributions are twofold. First, we propose a new economic index, NSI, for Korea, and comprehensively evaluate its validity and utility. Second, we provide a practical framework for analyzing Korean text data with machine learning approaches to compile an economic index. Our framework addresses various issues related to using text data as a source of a new economic indicator, such as building a Korean text classification model, adjusting imbalanced data, and standardization for stability control of the index. Furthermore, we provide a framework to improve the efficiency of public work for compiling public economic statistics automatically without human intervention. As an experimental attempt to compile regular economic statistics without surveying, NSI demonstrates that efficient observational studies can complement traditional surveying methods. The proposed process of automatic compilation of NSI can be applied to invent similar economic indices in different fields.

For future work, there are several interesting directions to explore using NSI. One potential avenue is to investigate the predictive power of the text-based index by incorporating it into existing economic forecasting or nowcasting models. Additionally, efforts to construct comparable text-based indices in different sectors and fields, such as production, employment, and inflation, would be beneficial. We hope that our work can inspire the use of machine learning techniques as innovative tools for economic research.

References

Armah, N. A. et al. (2013). Big data analysis: The next frontier. *Bank of Canada Review*, 2013(Summer):32–39.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Babii, A., Ghysels, E., and Striaukas, J. (2021). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, pages 1–23.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636.

Barsky, R. B. and Sims, E. R. (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review*, 102(4):1343–77.

Bennani, H. and Neuenkirch, M. (2017). The (home) bias of european central bankers: new evidence based on speeches. *Applied Economics*, 49(11):1114–1131.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2021). Business news and business cycles. Technical report, National Bureau of Economic Research.

Chollet, F. et al. (2015). Keras. https://keras.io.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Heaton, J. (2008). *Introduction to neural networks with Java.* Heaton Research, Inc.

Huang, C., Simpson, S., Ulybina, D., and Roitman, A. (2019). *News-based sentiment indicators.* International Monetary Fund.

Jeon, J.-J., An, S., Lee, M., and Hwang, H. (2020). 경제용어 감성사전 구축방안 연구 [study on the construction methods for the sentiment vocabulary of economic terminology]. *BOK 지역경제리뷰[BOK Regional Economics Review]*.

Kim, H., Im, J., Lee, H., and Lee, S. (2019). 온라인 뉴스 기사를 활용한 경제심리보조지수 개발 [development of economic sentiment supplementary index using online news articles]. *BOK 국민계정리뷰 [BOK National Accounts Review]*, 2019(2).

Kim, H.-j., Jeo, S.-r., and Kim, D. (2021). 경제 텍스트 데이터를 활용한 키워드 분석방안 연구 [study on keyward analysis methods using economic text data]. *BOK 국민계정리뷰 [BOK National Accounts Review]*, 2021(1).

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, Y. J., Kim, S., and Park, K. Y. (2019a). Deciphering monetary policy board minutes with text mining: The case of south korea. *The Korean Economic Review*, 35(2):471–511.

Lee, Y. J., Kim, S., and Park, K. Y. (2019b). Measuring monetary policy surprises using text mining: The case of korea. *Bank of Korea WP*, 11.

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.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moon, H. (2019). 빅데이터의 경제통계 활용 현황 및 시사점 [current status and implication of the use of big data in economic statistics]. *한국경제포럼[Korean Economic Forum]*, 11(4):89–105.

Nguyen, K., La Cava, G., et al. (2020). Start spreading the news: News sentiment and economic activity in australia. *Sydney: Reserve Bank of Australia*, 33.

Ryu, W. H. (2014). Open korean text (okt). https://github.com/open-korean-text/open-korean-text.

Seki, K., Ikuta, Y., and Matsubayashi, Y. (2022). News-based business sentiment and its properties as an economic index. *Information Processing & Management*, 59(2):102795.

Seo, B., Lee, Y., and Cho, H. (2022). 기계학습을 이용한 뉴스심리지수 (nsi)의 작성과 활용 [compilation and utilization of news sentiment index (nsi) using machine learning]. *BOK 국민계정리뷰 [BOK National Accounts Review]*, 2022(1).

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of econometrics*.

Thorsrud, L. A. (2016). Nowcasting using news topics. big data versus big bank. *Working Papers, Centre for Applied Macro and Petroleum economics (CAMP), BI Norwegian Business School*, 2016(6).

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

van Aarle, B. and Kappler, M. (2012). Economic sentiment shocks and fluctuations in economic activity in the euro area and the usa. *Intereconomics*, 47(1):44–51.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Webster, J. J. and Kit, C. (1992). Tokenization as the initial phase in nlp. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Won, J.-H., Son, W., and Moon, H. (2017). 텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류 [economic-sentiment-related article classification using text mining techniques. *BOK 국민계정리뷰 [BOK National Accounts Review]*, 2017(4).