



BOK Working Paper



Machine-Learning-Based
News Sentiment Index (NSI) of Korea

Beomseok Seo, Younghwan Lee, Hyungbae Cho



2022. 9



Economic Research Institute
Bank of Korea

Publisher
Changyong Rhee
(Governor of Bank of Korea)

Editor
Yang Su Park
(Director General of the Institute)

BOK Working Paper is occasionally published by the Economic Research Institute, Bank of Korea. This is circulated in order to stimulate discussion and comments. Articles include research achievement by the staff and visiting scholars, and selected works sponsored by the Institute.

The views expressed in this paper do not necessarily reflect those of Bank of Korea or the Economic Research Institute.

Requests for copies of publications, or for addition/changes to the mailing list, should be sent to:

Economic Research Institute
Bank of Korea
39 Namdaemunno Jung-Gu
Seoul, 110-794, Korea

E-mail: eso@bok.or.kr

Fax: 82-2-759-5410

This publication available on the BOK
Economic Research Institute website
(<http://imer.bok.or.kr>)

© Bank of Korea, 2022
All rights reserved.

Reproduction for educational and non-commercial purposes is permitted provided that the source is acknowledged.

Machine-Learning-Based News Sentiment Index (NSI) of Korea

Beomseok Seo^{*}, Younghwan Lee^{**}, Hyungbae Cho^{***}

The views expressed herein are those of the authors, and do not necessarily reflect the official views of the Bank of Korea. When reporting or citing this paper, the authors' names should always be explicitly stated.

^{*} Corresponding author: Economist, Department of Economic Statistics, Bank of Korea, Tel: +82-2-759-5253, 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 and Dr. Heejin Hwang and Dr. Ahrang Lee for helpful comments.

Contents

I. Introduction	1
II. Compiling News Sentiment Index (NSI) of Korea	4
III. Validity Assessments of NSI	15
IV. Utility Assessments of NSI	20
V. Summary and Discussion	24

Machine-Learning-Based News Sentiment Index (NSI) of Korea

We develop the Korean news sentiment index (NSI) that measures the economic sentiment of Korea by computing it daily from the news texts scrapped from the internet. We use a set of natural language processing techniques and develop a state-of-the-art transformer-neural-network-based sentiment classifier particularly designed for computing NSI of Korea. The proposed model handles large news samples effectively and computes NSI efficiently. NSI is more frequently and immediately compiled than official indices based on monthly surveys, and hence, helps to identify changes in economic sentiments before the official statistics are released. Also, NSI provides explanations for why the economic sentiments fluctuate via its keyword analysis and sector indices. NSI is designed to be compiled automatically. We assess the validity and utility of NSI from multiple perspectives. The assessments support our findings that NSI is useful as a leading index and informative to find inflection points in economic sentiments.

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

JEL Classification: C45, C82, E32

I. Introduction

Developing a frequently and immediately available economic indicator has been in the center of interests for economists and central bankers (Barsky and Sims, 2012). The traditional methods rely on surveys to investigate the economic sentiments by asking people how they feel about economy, while the recent advances in natural language processing (NLP) technologies have enabled the use of unstructured text data to predict the economic cycles (Armah et al., 2013; Bennani and Neuenkirch, 2017; Gentzkow et al., 2019; Athey and Imbens, 2019).

This paper propose a novel Korean news sentiment index (NSI) that measures the economic sentiment of Korean domestic economy by computing it on a daily basis using the news texts scrapped from the internet and a set of natural language processing (NLP) techniques. The proposed index is more frequently and immediately compiled than official economic sentiment indices which are based on monthly surveys, and hence, helpful to identify inflection points in economic sentiments before the official statistics are released. In this paper, we explain the overall process of building the NSI of Korea using machine learning approach, and provide in-depth assessments of the newly computed index to evaluate its validity and utility in the economics perspective.

News text data has received increasing attentions as it becomes considered as a new source of information (Gentzkow et al., 2019; Moon, 2019). News text data is resourceful because it includes a huge amount of information (volume) in various economic topics (variety) and spreads the information immediately (velocity). Many works to extract economic information from news texts have been studied in both academia and public organization in the world including central banks. News text data has been studied for the various purposes in economics: to measure economic uncertainty (Baker et al., 2016), to improve the forecasting accuracy of the economic cycle (Bybee et al., 2021; Seki et al., 2022), and to predict the inflation expectation (Larsen et al., 2021).

Especially, news sentiment index (NSI) is the topic that has been widely studied using news text data (Shapiro et al., 2020). NSI is an index that is computed by counting the positive and negative sentences in news articles and compiling the difference between two numbers as an economic index. Unlike traditional eco-

conomic sentiment indices relying on surveys, which are time and cost-consuming, NSI is based on news articles that are available from the internet, and hence it has the advantage of identifying economic sentiment more frequently and immediately than the traditional statistics with significantly lower cost. However, the process of computing NSI is not trivial. Unlike the survey-based sentiment statistics which are under a strict quality control process to ensure the stability of the index, extracting economic sentiments from news texts involves various tricky problems that should be resolved to use it as a reliable economic indicator. The first hurdle to use news text as a source of economic sentiment is that the unstructured text data may include more noise than the structured survey data. To mitigate the noise in text data, a large sample is preferred to compute a stable index. The second hurdle is how to decide the economic sentiment of the news articles consistently because the interpretation of news articles may differ according to the evaluators. Especially these hurdles become a bigger problem when the large samples of news articles are required to be evaluated by a few evaluators. Hence, it is impractical for human to classify all the news articles to evaluate the economic sentiments of news texts. On the contrary, using NLP and text mining technologies makes it more feasible to compute NSI since it provides a way to build a consistent classifier that can handle the large sample of news texts efficiently.

Developing NSI using NLP and text mining techniques has been mainly studied among the researchers in central banks and international organizations who need to judge the economic situation quickly. Especially, the researchers in central banks lead the initial development of NSI to introduce a timely sentiment index as the economic sentiments have a significant impact on the effect of the monetary policy of the central banks. Federal Reserve Bank of San Francisco releases daily news sentiment index (NSI) by analyzing news texts based on lexical approach (Shapiro et al., 2020). Lexical approach for sentiment classification classifies a sentence or an article into a positive or negative sentiment based on a pre-defined dictionary of specific words (Hamilton et al., 2016). International Monetary Funds (IMF) (Huang et al., 2019) and Reserve bank of Australia (Nguyen et al., 2020) have published similar works for NSI of different regions based on lexical approach. Central bank of Norway (Thorsrud, 2020)

uses unsupervised learning techniques to find news indices by topics. Also, in academia, news texts are being actively studied to increase the accuracy of economic forecasting (Thorsrud, 2016; Babii et al., 2021). Bank of Korea has also researched on text data to develop economic indices by developing lexical dictionary for economics (Jeon et al., 2020), using topic modeling (Won et al., 2017), and analyzing keywords in news articles (Kim et al., 2021). There has been an attempt to develop a supplementary sentiment index of Korea using online news data as well (Kim et al., 2019) but their work does not use the state-of-the-art NLP model to analyze Korean texts and does not carry out the in-depth analysis of the impact.

Unlike many other works using lexical approach for English text data, our proposed NSI is computed directly based on 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 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 simple explanation.

To use the machine learning approach on Korean news text data, we first construct the big training samples that consist of randomly chosen news sentences and corresponding labels classified by human. Then, we use the training samples to develop a precise classifier for NSI of Korea. To this end, we design a new model based on the transformer neural network classifier using its encoder structure for classification, and compute 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.

For validity assessments, we not only evaluate the classification accuracy of the new classifier but also investigate if the newly proposed index can reflect the

true economic cycle and do so even prior to the official statistics. For this purpose, we conduct the comparison analysis of the computed monthly NSI with other economic sentiment indices and real economic indices, and also, we 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 biggest benefits of computing NSI from news texts is that it is easy to investigate the reasons why the index fluctuates via its keyword analysis and sector indices. In addition, we also investigate the temporal priority of daily NSI by reviewing the cases when the NSI reacts prior to official statistics.

This paper provides contributions in three aspects: 1) We provide the NSI of Korea as a regular statistics with an in-depth analysis of its validity and utility. The computed daily and monthly NSIs are now publicly available through Economic Statistics System (ECOS) in Bank of Korea (ecos.bok.or.kr). The NSI has been registered as an experimental statistics in Korea (No. 2022-001), which is a concept introduced by Statistics Korea to promote the use of big data for public statistics. 2) We provide a practical framework of analyzing Korean text data with machine learning models to compile an economic index. 3) We provide a framework for compiling public statistics by automation without human intervention, which increases the efficiency of public work for compiling statistics. This framework may also be helpful for developing new indices in other sectors.

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

II. Compiling News Sentiment Index (NSI) of Korea

1. The Concept of News Sentiment Index

News sentiment index (NSI) is computed by counting the positive and negative sentences of daily news articles. Hence, classifying the sentiment of the vast

amount of news articles efficiently with high accuracy is the core of computing precise NSI. To this end, we use machine learning approach. The machine learning approach analyzes the patterns of news sentences that are classified by humans to train a statistical model, and then the model is applied to a new sentence to classify its sentiment. The key of machine learning approach is to build the following sentiment classifier that fits the data pairs, $\{(s, l)_i\}_{i=1}^N$, well for a certain news sentence, $s_i \in \mathcal{S}$, and its sentiment label classified by humans, $l_i = (p_i^{(0)}, p_i^{(1)}, p_i^{(2)})$, where $p_i^{(0)} = P(s_i \text{ is positive})$, $p_i^{(1)} = P(s_i \text{ is negative})$, $p_i^{(2)} = P(s_i \text{ is neutral})$. That is, for the model parameter vector, θ , the sentiment classifier f is as follows.

$$f_{\theta} : \mathcal{S} \rightarrow \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}. \quad (1)$$

2. News Text Data

As the input of the model (1), we first explain the news text data and its source. Text data of news articles are collected through the web scrapping technique. Web scrapping is a method directly downloading publicly available data from the internet. We collect news articles by applying the web scrapping technique on an internet news portal. The portal site classifies news articles into six categories based on the choice of news article authors: politics, economy, society, life/culture, IT/technology, and world. For NSI of Korea, we collect the news articles from only the economy section to restrict the scope of the news sentiment we compute into the economic sentiment. The entire news articles released from 2005 on the news portal are collected. The constructed news database consists of about 50 media companies' news articles, including TV broadcasters, internet news agencies, and regional news magazines. The average number of daily news articles on weekdays is about 4,000 as of 2021. We cannot find the official statistics of the total number of news articles created in Korea, but we suppose the number of news we gather is quite close to the total number considering more than 50 major media companies post their news on the portal site.

When the news articles are collected via web scrapping, data quality can be an issue because the web scraper collects duplicate data and advertising articles as well. To handle this issue, if a newly collected article is the exact duplicate of

any news in database within 30 days, we exclude the article without adding it into the database.

3. Preprocessing for News Text

In order to apply machine learning on the news text data, it is important to preprocess the text data to convert it into numeric that the sentiment classifier of (1) can understand. The preprocessing steps include establishing data structure, tokenizing the structured text data using part-of-speech (POS) units (Webster and Kit, 1992), and converting the tokenized text data into numeric data. These processes are conducted through sequential applications of text mining techniques.

First, it is necessary to determine the input structure of the sentiment analysis, i.e., whether to classify the sentiments based on articles or sentences. It is common that a news article addresses both positive and negative sentiments. Therefore, it is not effective to classify the sentiment of a news article based on the entire texts of the article. To avoid this problem, NSI is computed based on the randomly sampled sentences from news database rather than using the entire articles. That is, the input data is each sentence $\{(s, l)_i\}_{i=1}^N$ randomly selected from news database.

Second, to make a sentiment classifier understand text data, it is necessary to tokenize the texts in a sentence into the units of POS which is the smallest unit of a word assigned in accordance with its syntactic functions in Korean. This process is called POS tokenization. For example, a korean sentence ‘뉴스심리지수를 작성하였다.’ is tokenized as follows.

‘뉴스심리지수’(proper noun) + ‘-를’(object case marker) +
 ‘작성’(general noun) + ‘하’(verb derivative suffix) + ‘-았’(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 statis-

tical 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.

$$g : S \rightarrow T. \quad (2)$$

The choice of tokenizer is language specific and dependent on which fields the tokenizer is used for because the tokenizer decides whether a compound word should be split into multiple root words with separate meanings or interpreted as one word. The results of tokenization have significant effects on the final classification performance. In our work, we use a pretrained open Korea text (OKT) tokenizer for g , which is available as an open source. Our text data to measure economic sentiments includes various fields and general expressions, and has a huge volume. Hence, we choose the OKT tokenizer, which is widely used to analyze Korean texts for the general purpose and shows high efficiency in terms of computational time.

Lastly, the tokenized data is transformed into numerical digits by integer encoding. The integer encoding matches an integer to each distinct corpus token, and represents an input sentence as a numeric vector. The maximum length, m , of the numeric vector is limited to 80 for each sentence for a unified input data structure. That is, for the sentence with its length less than 80, zeros are padded in front of the numeric vector, and for the sentence longer than 80 tokens, its last tokens are truncated. The maximum length, $m = 80$, is chosen according to 99 percentile of sentence lengths in news database. The integer encoding assigns a sequence of tokens, $\tau_i \in T$, to an integer vector, $x_i \in \mathbb{R}^m$, by the following bijection function h .

$$h : T \rightarrow \mathbb{R}^m \quad (3)$$

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.

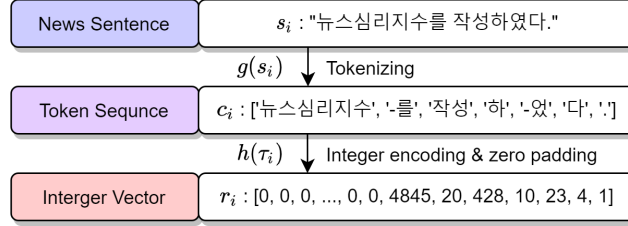


Figure 1. Preprocessing steps of Korean news texts to compute NSI of Korea.

4. Construction of Korean News Sentiment (KoNS) Classifier

Specifically for the sentiment classifier (1), we build a classifier based on the state-of-the-art transformer neural network. Recently, the transformer model is widely used for various natural language processing tasks. The transformer consists of a multiple-head attention and a feed-forward network, and is known to have advantages in perceiving context of a sentence (Vaswani et al., 2017) due to its multiple-head attention structure. The attention structure refers to an artificial neural network structure in a sequential model that is configured to give a higher weight to the inputs that need to be learned more intensively than others. We build the sentiment classifier for NSI of Korea, which we call Korean News Sentiment (KoNS) classifier, using the encoder structure of the transformer model.

Figure 2 is a schematic diagram of the KoNS classifier established to compute NSI of Korea. KoNS is designed based on the following multihead attention structure in a transformer block that is introduced by Vaswani et al. (2017).

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \quad (4)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (5)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (6)$$

where d_k is the dimension of queries, Q , and keys, K ; d_v is the dimension of values, V ; and h is the number of heads. For the output dimension, 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 $W_i^O \in \mathbb{R}^{hd_v \times d_m}$ are the model parameters.

Since KoNS is a classification model, we use self-attention mechanism, i.e., the preprocessed input sequence x_i is used for all Q , K , and V . We set $d_k = d_v = d_m = 32$ and $h = 2$. The output of the multihead attention is attached to the original preprocessed input, and followed by a layer-normalization (Ba et al., 2016) which is introduced 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.

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

where $W^F \in \mathbb{R}^{d_m \times d_f}$, $b^F \in \mathbb{R}^{d_f}$, and $\sigma(\cdot)$ is a rectified linear unit (ReLU) activation function which is equivalent to the elementwise max function. The x in (7) indicates the output of the previous layer. Through the preprocessing and the transformer block, a news sentence is transformed into a numerical matrix of 80×32 dimensions. Subsequently, a feed-forward network of dimension 3 with a followed softmax function generates the predicted probability of sentiments: positive, negative, and neutral. For the more detailed aspects of the transformer neural network used to build KoNS, we refer the reader to Vaswani et al. (2017).

The choice of the configuration of the transformer neural network including its hidden unit dimensions, d_k, d_v, d_m, d_f , is made following the conventions used for the transformer model for similar text classification tasks (Chollet et al., 2015). It is known that in practice, a large enough number of hidden layers and units are required to express a complex function (LeCun et al., 2015) nonetheless there is no theoretical reason to use more than two layers (Heaton, 2008). A general rule of thumb is using a comparable configuration for similar tasks and data.

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

$$\hat{l}_i \equiv \hat{f}_\theta(s_i) \quad (8)$$

$$= \text{TransformerBasedClassifier} \circ h \circ g(s_i) \quad (9)$$

where $\hat{l}_i \in \{(\hat{p}_i^{(0)}, \hat{p}_i^{(1)}, \hat{p}_i^{(2)})\}$, and $0 \leq \hat{p}_i^{(c)} \leq 1$ for $c = 0, 1, 2$, and $\sum_{c=0}^2 \hat{p}_i^{(c)} = 1$.

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, $s_i, i = 1, \dots, N$.

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

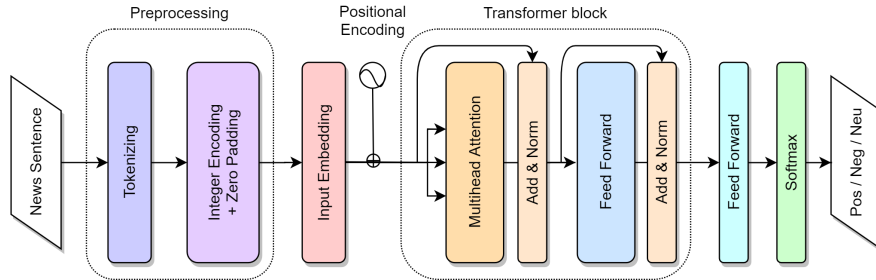


Figure 2. A schematic diagram of the transformer-based sentiment classifier for NSI of Korea.

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.

5. Training Data and Model Training

To estimate θ in KoNS correctly, it is important to construct the training data with correct labels. We construct the training data with 446,478 sentences chosen across 2005 to 2021 based on a naively designed stratified sampling by

year. The training data is constructed in three times between 2019 and 2021. In the first attempt in 2019, we created 230,583 labeled news sentences that are sampled from the news generated between 2008 and 2018, and we added more sentences to the training data later using 84,069 sentences based on the news generated between 2008 and May 2020 at the second attempt, and at the last attempt, 131,826 sentences generated between 2005 and June 2021. Although the training samples are not chosen in a rigorous manner, we conclude our training data does not cause any critical problem for news text sentiment classification since text expressions revealing sentiments would not change dramatically in a decade, especially if they are for news sentences.

The training sentences are classified into one of the positive, negative, and neutral sentiments by 16 trained personnel in total. Sentiment labels are reviewed by other reviewers after initial classification to reduce the measurement errors. However, even for humans, there are confusing sentences of which sentiments are difficult to classify. Equal sentences can transmit different tones according to their subjects and contexts. For instance, inflation may be detrimental to the economy, particularly for consumers, while it may also be beneficial to property owners. That is, the change in inflation cannot be classified as either positive or negative. To handle this problem, we establish minimal guidelines for the frequent subjects in economic news as in Table 1.

The guidelines in Table 1 impose limitations on sentiment interpretation and result in a more conservative classification by making more clear tones represented in the training data. As a result, about 80% of the total sentences in the training data are classified neutral. The half of the remaining 20% are classified positive, and the other half are negative. Because the economic sentiment is indeed abstract, the subjectivity in sentiment interpretation is unavoidable when building training data for the use of machine learning approach. The subjectivity is inherent also for the traditional surveying approach because each person uses a distinct set of criteria to assess their economic situation.

Meanwhile, the imbalanced class weights of the training data have a significant influence on the final prediction of a classifier. Taking into account the predicted accuracy of KoNS and the fluctuation of the computed NSI, we adjust

Subjects	Guidelines
Stock price, interest rates, exchange rates, and other asset prices	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Advertisements for apartment sales are neutral. • Statements on the unsold apartments are classified as positive or negative.
Government and public sectors	<ul style="list-style-type: none"> • 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.

the loss function in (10) with class weights, $w^{(c)} = \frac{N}{3N^{(c)}}$, as follows, where $N^{(c)}$ is the number of sentences in the class c of the training data.

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

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

6. Prediction for Daily News Data

Finally, the NSI is compiled by applying KoNS classifier onto the daily news data. As the data source for compiling daily NSI, the number of news articles we collect from the internet is about 4,000 per day for weekdays as of 2021, which is equivalent to about 70,000 sentences. The number of sentences drops to about 10,000 sentences per day for weekends. Therefore, to reduce the computing cost and make the NSI stable across days, we randomly sample 10,000 sentences every day and use only the sampled sentences to compute NSI. This sampling process prevents certain days from dominating the economic sentiment of NSI as more news articles are released during that days.

7. Computing News Sentiment Index (NSI) of Korea

Specifically, after we obtain the predicted sentiments of daily news sentences using KoNS, we compute the daily NSI as follows. First, we count the numbers of positive and negative sentences from the sample sentences that are collected during the previous 7 days of a particular day. Then, the daily NSI is computed as the ratio between the difference and the sum of the positive and negative counts. Here, we use 7 days of news sentences to generate a smooth and stable index. Monthly NSI is computed by the same procedure, but this time based on the news sentences collected between the 1st of the month and the closest previous Sunday, and the number is updated every Tuesday until the end of the month when the monthly NSI is finally computed by the news sentences of the month.

Both daily and monthly NSIs are standardized using their long term averages

to make them easy to be compared to the past economic sentiment. NSIs are standardized so that the averages and standard deviations are equal to 100 and 10 respectively. The starting time point of the standardization interval is fixed to 2005, and the last time point is extended to the end of the previous year at the beginning of every year. Specifically, the daily NSI of a particular day t , $NSI_t^{(\text{daily})}$, is compiled by the following formula.

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

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

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

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.

$$X_v = \frac{\sum_{v \in M_v} P_v - \sum_{v \in M_v} N_v}{\sum_{v \in M_v} P_v + \sum_{v \in M_v} N_v}, \quad (15)$$

where M_v is the index set of days belonging to the particular month v .

If NSI is greater than 100, it means that the economic sentiments in news articles are more optimistic than the past average, and more pessimistic, on the other hand, if it is less than 100.

8. Automated Compilation

NSI is designed to be compiled every week automatically without human labor. That is, the entire processes of web-scraping of daily news articles from the internet, text preprocessing, sentiment prediction, and compilation of the daily and monthly NSIs are carried out through a python script runs by an automated batch. The automated process collects news articles from the internet from the last point of the database to the previous day at 5 AM every day. Then, KoNS subsequently operates to compute the daily and monthly NSIs. The automation of the entire compiling process improves the efficiency of statistics compilation and saves labors and costs compared to the survey approach.

III. Validity Assessments of NSI

In this section, we validate 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, and impulse response analysis using a VAR macroeconomic model.

1. Sentiment Classification Accuracy

The accuracy of KoNS classifier is compared to other statistical models based on a small validation dataset. The validation data is constructed by 5,000 randomly sampled sentences from the news database in 2021 as a out-of-sample dataset, and labels are created by an experienced personnel working in Bank of Korea who is assumed to know the true economic implication of news. The validation data consists of 7% positive, 7% negative and 86% neutral sentences. The difference of the class proportions between the training and validation data can cause a decrease in validation accuracy and may generate a consistent bias in NSI. Nevertheless, the systematic bias is mitigated by the standardization process in compiling NSI, and considering the fact that the proportion fluctu-

ates according to the economic situation, we conclude that using the validation dataset to evaluate KoNS is not problematic.

Table 2 shows the classification accuracy of KoNS and other standard models. We compare only the positive and negative classes because the neutral class does not affect the computation of NSI. Table 2 indicates that KoNS with weight-adjusted loss has the best performance. Meanwhile, one interesting finding is that the simple linear model 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.

	LR	FFN	SVM		KoNS	
	SL	SL	SL	WAL	SL	WAL
Accuracy	0.90	0.88	0.86	0.86	0.96	0.98
Sensitivity	0.97	0.94	0.87	0.91	0.95	0.98
Specificity	0.82	0.78	0.84	0.79	0.97	0.99
Precision	0.86	0.87	0.86	0.88	0.95	0.97
F-1 score	0.92	0.90	0.86	0.89	0.96	0.98

Table 2. The binary classification accuracy comparisons for positive and negative classes, i.e., the neutral class is ignored in both predicted and true labels. LR: logistic regression, FFN: feed-forward network, SVM: support vector machine, SL: standard loss used, WAL: weight-adjusted loss used.

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

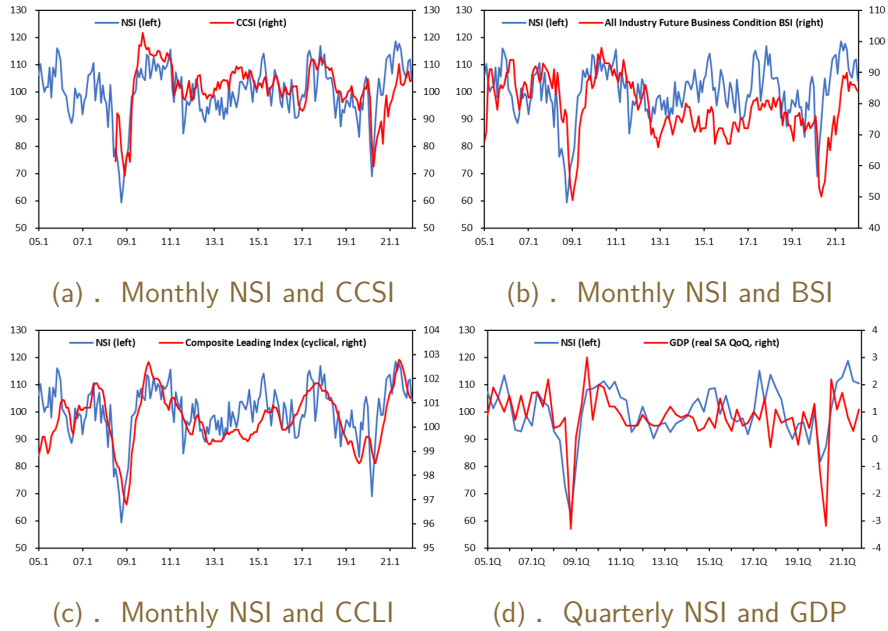


Figure 3. Comparisons between NSI and various economic indicators.

Economic indicators		Max. Corr.	Max. Corr. Lag
CCSI		0.75	-1
ESI		0.61	-2
CSI	Living Standard of Household	0.74	-1
	Domestic Economic Situation	0.73	-1
	Expectation of Living Standard	0.73	-1
	Expectation of Domestic Economic Situation	0.70	-1
	Expectation of Household Income	0.68	-1
	Spending Plan	0.57	-1
	Expectation of Employment Situation	0.72	-1
BSI	All Industries Business Condition	0.64	-1
	All Industries Profitability	0.68	-1
	All Industries Financial Situation	0.64	-1
	All Industries Future Business Condition	0.61	-2
	All Industries Future Profitability	0.65	-2
	All Industries Future Financial Situation	0.61	-2
Real Indices	KOSPI(monthly closing price YoY%)	0.68	-1
	Cycle of Composite Leading Index(CCLI)	0.76	-2
	GDP (real SA QoQ%)	0.53	0

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

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 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. 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.

$$Cy_t = \alpha + \sum_{i=1}^k C_i y_{t-1} + \varepsilon_t, \quad (16)$$

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

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 $e_t = C^{-1}\varepsilon_t$, which indicates C^{-1} has the lower triangular structure.

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

Table 4. 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.

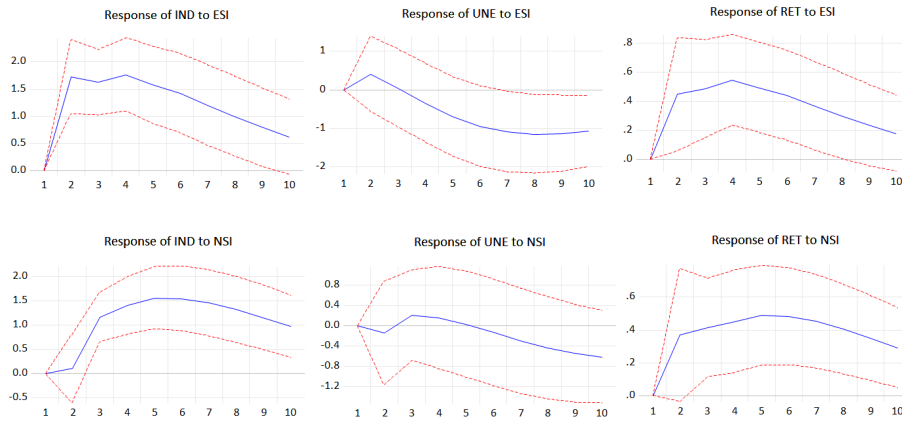


Figure 4. Impulse responses of macroeconomic variables on the economic sentiment shocks measured by ESI and NSI.

We recover the orthogonalized shocks with $\varepsilon = Ce_t$ 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 4 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 R^2 close to that of ESI model. The impulse responses of the two models are displayed in Figure

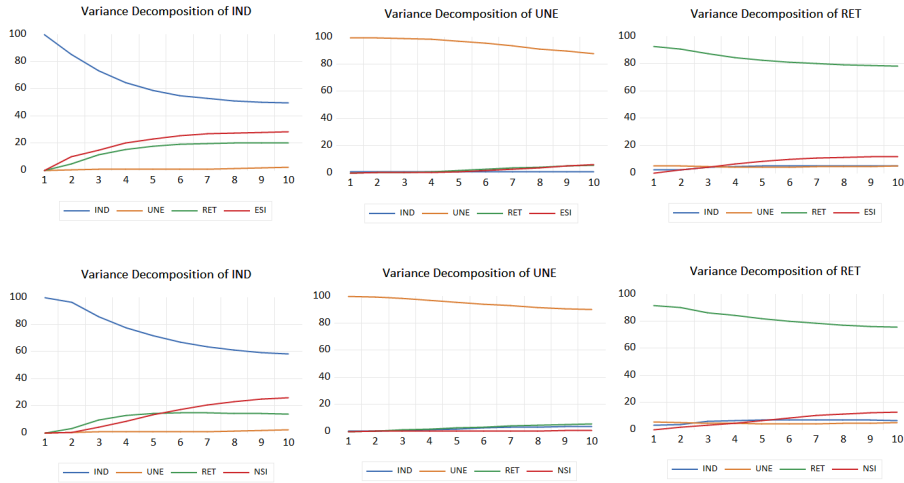


Figure 5. Variance error decomposition for the VAR models with ESI and NSI. The sentiment shocks measured by ESI or NSI largely contribute to the variance in the business cycle variable, IND.

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 biggest 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.

IV. Utility Assessments of NSI

In this section, we provide utility assessments of NSI. The biggest utility of NSI is that the daily NSI is a timely economic indicator. We examine the temporal priority of daily NSI by comparing the ability to find inflection points of economic sentiments to other official statistics. Another big benefit of NSI

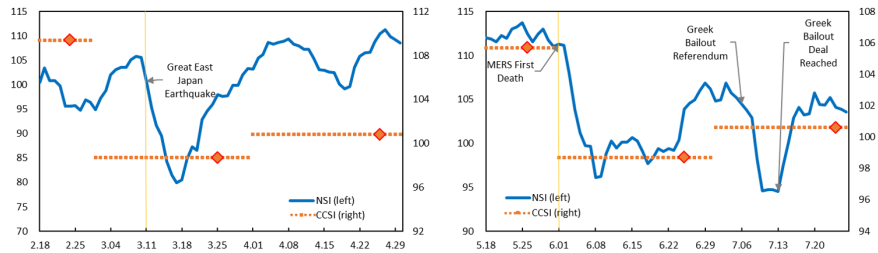
is that it is self-explainable by providing more detailed information about its probable factors of fluctuation via keywords and metadata. We examine NSI using keywords analysis and sector NSIs computed for three economic sectors.

1. Temporal Priority

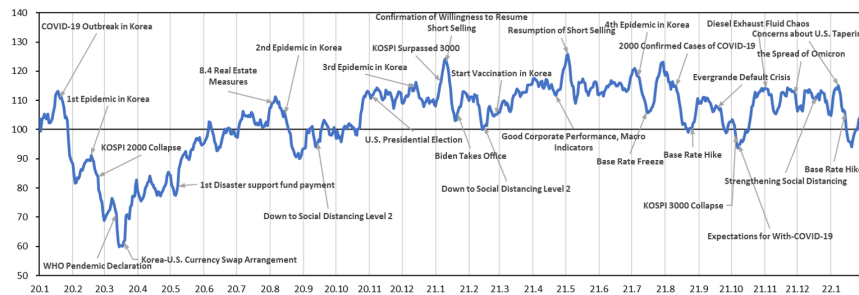
Because NSI is compiled on a daily basis, it has the advantage of quickly identifying the changes in economic sentiments prior to the official statistics that are released monthly based on surveys. As illustrated in Figure 6, NSI can immediately quantify the impact of important issues in economy and detect the inflection point in economic sentiments effectively. 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 of CCSI was observed at the end of March when the statistics was released after the monthly survey was made. Also, in 2015, NSI dropped immediately after the first death of MERS was known to the public on June 1st, but CCSI could not indicate the event until the end of June. That is, NSI provides immediate information earlier than survey-based official statistics and can be used as a supplementary indicator to detect the inflection point of economic sentiments quickly.

2. Explainability and Keyword Analysis

Another big advantage of NSI is that the analysis of keywords reveals the reason why NSI fluctuates. Figure 7 shows the keyword networks in positive and negative news articles which are classified by KoNS in the last week of November 2021. While the positive keywords of the period mainly consist of company earnings, the negative keywords are related to the price hike of raw materials and COVID-19 Omikron mutation. The keywords can also reveal the main economic issues of a time. In the last week of November 2021, ‘Omikron’ mutation, ‘KOSPI’, and the government’s deliberation on the ‘budget’ for supporting the small business owners have been mentioned more compared to the past week, which indicates the week’s main economic issues. Table 5 shows the keywords of the same week in positive and negative news articles divided by sub-sectors: macro, finance, and industry. The table demonstrates that the impact of the



(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.

stock market fluctuations was mentioned in both positive and negative articles in the finance sector, while in the macro sector, the rise of exports in November was largely responsible for the rise of NSI but the worsening industrial production statistics was attributed as a factor for the decline in NSI. These detailed information in economic sentiments is hard to be obtained through surveys.

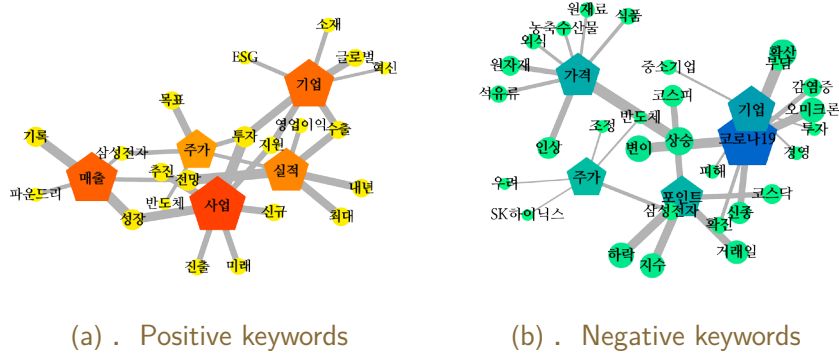


Figure 7. The keyword networks of positive and negative sentences classified by KoNS. The pentagon nodes indicate the 5 biggest main keywords appearing in the sentences in the last week of November 2021, and the circle nodes indicate the related keywords appearing together with the main keywords.

3. Sector NSIs

NSI can be computed by sectors easily by dividing the news sources into sectors. This is an advantage of using news articles to compute economic sentiment. The sector information of the news articles is obtained without efforts as metadata when we collect the news articles. Internet portals provide news in categories, and hence, we can use the categories to divide news sources. By dividing the news articles into three sub-categories such as macro, finance, and industry, we compute sector NSIs. As shown in Figure 8, although the sector NSIs have high correlations with the aggregated NSI, they show different patterns in particular time points when specific events occur such as the COVID-19 outbreak and company earning announcements. Sector NSIs provide more detailed information on economic sentiments in different sectors. Table 6 shows that the sector NSIs become more homogeneous after the COVID-19 outbreak, which imply the

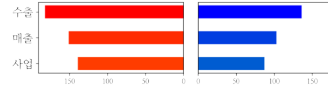
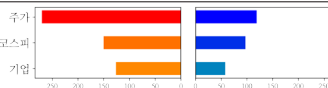

Sec.	Positive sentences		Negative sentences	
	Related words	Keywords	Keywords	Related words
Mac.	(증가, 11월, 최고, 실적) (분기, 증가, 기록, 성장) (투자, 지원, 글로벌, 성장)		(확산, 변이, 오미크론, 위기) (상승, 석유류, 외식, 원자재) (감소, 자동차, 광공업, 반도체)	
Fin.	(상승, 전망, 실적, 삼성전자) (지수, 외국인, 상승, 코스닥) (성장, 투자, 확대, 미래)		(하락, 지수, 오미크론, 최저) (하락, 삼성전자, 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 5. 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.

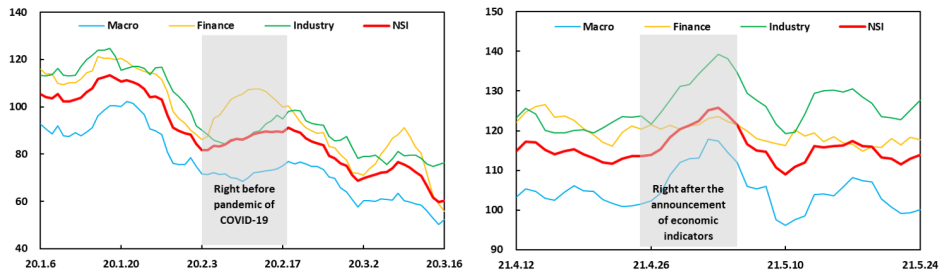
hypothesis that the COVID-19 dominated the factors in economic sentiment fluctuations after its outbreak. To obtain such information through survey, the cost would increase significantly by adding more questions.

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

V. Summary and Discussion

In this paper, we compute a news sentiment index (NSI) of Korea based on the news articles collected from the internet using web-scraping techniques. To compute NSI, we develop a Korean news sentiment (KoNS) classifier based on the state-of-the-art natural language processing (NLP) model. We build KoNS using the encoder structure of the transformer model and train the model with



(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.

a modified loss function so that the model can control the imbalanced data. KoNS is trained with more than 450 thousand training sentences that are labeled by 16 trained personnel. Finally, daily and monthly NSIs are computed based on the counts of the positive and negative sentences classified by KoNS for the daily news sentence samples. The validity of the computed NSI of Korea is evaluated in multiple perspectives. Cross-correlation analysis shows that NSI leads composite consumer sentiment index (CCSI) based on a monthly survey by 1 months with its correlation equal to 0.75 and cycle of composite leading index (CCLI) by 2 month with its correlation of 0.76. Also, the impulse response analysis demonstrates that a rise in NSI stimulates the industrial production with its hike appearing in 3 months. We, also, evaluate the various utilities of NSI. Daily NSI shows that it correctly detects inflection points in economic sentiments prior to official statistics based on monthly surveys. Keyword analysis reveals the factors why NSI fluctuates and provides more information that cannot be quantified through survey-based statistics. Lastly, sector NSIs are easily computed by dividing news articles into sectors and they addresses more detailed information for economic sentiments in sectors.

This paper has contributions in three folds. First, we propose a new economic index, the news sentiment index (NSI) of Korea, with in-depth analysis of validity and utility assessments of the newly computed index. Second, we provide a practical framework for analyzing Korean text data with machine learning approach to compile an economic index. We handle various issues encountered to use 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. Lastly, we provide a framework to improve the efficiency of the public work for compiling public economic statistics automatically without human intervention. As an experimental attempt to compile the regular economic statistics without surveying, the NSI demonstrates that the efficient observational study can complement the traditional surveying method. The proposed process of automatic compilation of NSI can be applied in inventing similar economic indices in different fields.

For the future work, there are various interesting directions to use NSI. One can investigate the forecasting power of the text-based index by adding it into an existing economic forecasting model or a nowcasting model. Also, it must be helpful effort to construct comparable text-based indices in different sectors and fields, such as production, employment, and inflation. We hope our work can encourage the use of machine learning techniques as new 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.
- Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, volume 2016, page 595. NIH Public Access.

- 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 경제심리 보조지수 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 keyword 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.
- 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.

- 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).

〈Abstract in Korean〉

한국의 기계학습 기반 뉴스심리지수

서범석*, 이영환**, 조형배***

본 논문은 인터넷에서 스크랩한 뉴스 기사를 일별로 분석하여 국내 경제주체들의 경제심리를 추정하는 한국의 뉴스심리지수(news sentiment index, NSI) 개발 방법을 제시하였다. 이를 위해 일련의 자연어처리 기법들을 활용하였으며 NSI 작성에 적합하도록 트랜스포머(transformer) 인공지능망 모델을 기반으로 감성분류 모델을 구축하였다. NSI는 월별 서베이에 의존하는 공식 통계보다 고빈도로 신속하게 작성하는 것이 가능하며 따라서 공식통계 발표 전에 경제심리 변화를 포착하는 데 유용하다. 또한, NSI는 경제심리 변화 요인을 키워드 분석과 부문별 지수 작성을 통해 파악 가능한 점도 장점이다. NSI는 사람의 개입 없이 자동으로 추계되도록 설계되었다. 본 논문은 작성한 NSI의 타당성과 유용성을 여러 각도에서 평가하였다. 평가 결과는 NSI가 선행 지표로서 유용하며 경제심리의 변곡점 포착에 유용한 정보를 제공할 수 있음을 시사한다.

핵심 주제어: 뉴스 텍스트 데이터, 경제분석을 위한 자연어처리, 심리 충격

JEL Classification: C45, C82, E32

* 한국은행 경제통계국 통계연구반 과장 (전화: 02-759-5253, E-mail: bsseo@bok.or.kr)

** 한국은행 경제통계국 통계연구반 과장 (E-mail: yhlee@bok.or.kr)

*** 한국은행 경제통계국 통계연구반 조사역 (E-mail: hyungbae.cho@bok.or.kr)

논고 작성에 많은 도움을 주신 문혜정 전 통계연구반 반장, 황희진 통계조사팀 팀장, 이아랑 거시경제연구실 차장께 감사의 말씀을 전합니다.

이 연구내용은 집필자 개인의견이며 한국은행의 공식견해와는 무관합니다. 따라서 본 논문의 내용을 보도하거나 인용할 경우에는 집필자명을 반드시 명시하여 주시기 바랍니다.

BOK 경제연구 발간목록

한국은행 경제연구원에서는 Working Paper인 『BOK 경제연구』를 수시로 발간하고 있습니다. 『BOK 경제연구』는 주요 경제 현상 및 정책 효과에 대한 직관적 설명 뿐 아니라 깊이 있는 이론 또는 실증 분석을 제공함으로써 엄밀한 논증에 초점을 두는 학술논문 형태의 연구이며 한국은행 직원 및 한국은행 연구용역사업의 연구 결과물이 수록되고 있습니다. 『BOK 경제연구』는 한국은행 경제연구원 홈페이지(<http://imer.bok.or.kr>)에서 다운로드하여 보실 수 있습니다.

- | | | |
|---------|--|---|
| 제2019-1 | Deciphering Monetary Policy Board Minutes through Text Mining Approach: The Case of Korea | Ki Young Park ·
Youngjoon Lee ·
Soohyon Kim |
| 2 | The Impacts of Macroeconomic News Announcements on Intraday Implied Volatility | Jieun Lee ·
Doojin Ryu |
| 3 | Taking a Bigger Slice of the Global Value Chain Pie: An Industry-level Analysis | Chong-Sup Kim ·
Seungho Lee ·
Jihyun Eum |
| 4 | Trend Growth Shocks and Asset Prices | Nam Gang Lee |
| 5 | Uncertainty, Attention Allocation and Monetary Policy Asymmetry | Kwangyong Park |
| 6 | Central Bank Digital Currency and Financial Stability | Young Sik Kim ·
Ohik Kwon |
| 7 | 은행의 수익 및 자산구조를 반영한 통화정책 위험선호경로 | 김의진 · 정호성 |
| 8 | 혁신기업에 대한 산업금융 지원: 이론모형 분석 | 강경훈 · 양준구 |
| 9 | 가계부채 제약하의 통화정책: 2주체 거시모형(TANK)에서의 정량적 분석 | 정용승 · 송승주 |
| 10 | Alchemy of Financial Innovation: Securitization, Liquidity and Optimal Monetary Policy | Jungu Yang |
| 11 | Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea | Youngjoon Lee ·
Soohyon Kim ·
Ki Young Park |
| 12 | Tracking Uncertainty through the Relative Sentiment Shift Series | Seohyun Lee ·
Rickard Nyman |
| 13 | Intra-firm and Arm's Length Trade during the Global Financial Crisis: Evidence from Korean Manufacturing Firms | Moon Jung Choi ·
Ji Hyun Eum |

14	특허자료를 이용한 우리나라 지식전파의 지역화 분석	이지홍 · 남윤미
15	Overhead Labour and Skill-Biased Technological Change: The Role of Product Diversification	Choong Hyun Nam
16	Does the Number of Countries in an International Business Cycle Model Matter?	Myunghyun Kim
17	High-Frequency Credit Spread Information and Macroeconomic Forecast Revision	Bruno Deschamps · Christos Ioannidis · Kook Ka
18	경제 분석을 위한 텍스트 마이닝	김수현 · 이영준 · 신진영 · 박기영
19	Takeover, Distress, and Equity Issuance: Evidence from Korea	Euna Cho
20	The Cash-Flow Channel of Monetary Policy: Evidence from Mortgage Borrowers	Sang-yoon Song
21	부의 효과의 분위 추정: 분위 정준 공적분 회귀를 중심으로	김기호
22	Identifying Government Spending Shocks and Multipliers in Korea	Kwangyong Park · Eun Kyung Lee
23	Systemic Risk of the Consumer Credit Network across Financial Institutions	Hyun Hak Kim · Hosung Jung
24	Impact of Chinese Renminbi on Korean Exports: Does Quality Matter?	Jihyun Eum
25	Uncertainty, Credit and Investment: Evidence from Firm-Bank Matched Data	Youngju Kim · Seohyun Lee · Hyunjoon Lim
26	A Structural Change in the Trend and Cycle in Korea	Nam Gang Lee · Byoung Hoon Seok

제2020 -1	인구 고령화가 실질 금리에 미치는 영향	권오익 · 김명현
2	달러라이제이션이 확산된 북한경제에서 보 유외화 감소가 물가 · 환율에 미치는 영향	문성민 · 김병기
3	상태공간 벡터오차수정모형을 이용한 월별 GDP 추정: 깃스표본추출 접근	김기호
4	우리나라 외환시장 오퍼레이션의 행태 및 환율변동성 완화 효과	박준서 · 최경욱
5	Common Factor Augmented Forecasting Models for the US Dollar–Korean Won Exchange Rate	Hyeongwoo Kim · Soohyon Kim
6	북한 「경제연구」로 분석한 경제정책 변화: 텍스트 마이닝 접근법	김수현 · 손 옥
7	북한의 광물 수출과 품목별 수입: 대중무역을 중심으로	김병연 · 김민정 · 김다울
8	Network–Based Measures of Systemic Risk in Korea	Jaewon Choi · Jieun Lee
9	Aggregate Productivity Growth and Firm Dynamics in Korean Manufacturing 2007–2017	Kyoo il Kim · Jin Ho Park
10	2001년 이후 한국의 노동생산성 성장과 인적자본: 교육의 질적 개선 효과를 중심으로	유혜미
11	House Prices and Household Consumption in Korea	Seungyoon Lee
12	글로벌 가치사슬 변화가 경제성장에 미치는 영향: 2008년 금융위기 전후 전 · 후방참여 효과의 국제비교를 중심으로	김세완 · 최문정
13	산업구조조정이 고용 및 성장에 미치는 영향	서병선 · 김태경
14	Cross–border Trade Credit and Trade Flows During the Global Financial Crisis	Moon Jung Choi · Sangyeon Hwang · Hyejoon Im

-
- | | | |
|----|--|--|
| 15 | International Co-movements and Determinants of Public Debt | Hasan Isomitdinov · Vladimir Arčabić · Junsoo Lee · Youngjin Yun |
| 16 | 북한 비공식금융 실태조사 및 분석 · 평가 | 이주영 · 문성민 |
| 17 | 북한의 장기 경제성장률 추정: 1956~1989년 | 조태형 · 김민정 |
| 18 | Macroeconomic and Financial Market Analyses and Predictions through Deep Learning | Soohyon Kim |
| 19 | 제조업의 수출과 생산성 간 관계 분석: 사업체 자료 이용 | 이윤수 · 김원혁 · 박진호 |
| 20 | 우리나라 제조업 수출기업의 내수전환 결정요인 분석 | 남윤미 · 최문정 |
| 21 | A Model of Satisficing Behaviour | Rajiv Sarin · Hyun Chang Yi |
| 22 | Vulnerable Growth: A Revisit | Nam Gang Lee |
| 23 | Credit Market Frictions and Coessentiality of Money and Credit | Ohik Kwon · Manjong Lee |
| 24 | 북한의 자본스톡 추정 및 시사점 | 표학길 · 조태형 · 김민정 |
| 25 | The Economic Costs of Diplomatic Conflict | Hyejin Kim · Jungmin Lee |
| 26 | Central Bank Digital Currency, Tax Evasion, Inflation Tax, and Central Bank Independence | Ohik Kwon · Seungduck Lee · Jaevin Park |
| 27 | Consumption Dynamics and a Home Purchase | Dongjae Jung |
| 28 | 자본유입과 물가상승률 간의 동태적 상관관계 분석: 아시아의 8개국 소규모 개방경제를 중심으로 | 최영준 · 손종철 |
-

29	The Excess Sensitivity of Long-term Interest rates and Central Bank Credibility	Kwangyong Park
30	Wage and Employment Effects of Immigration: Evidence from Korea	Hyejin Kim
제2021-1	외국인력 생산성 제고 방안—직업훈련 프로그램의 노동시장 성과 분석을 중심으로	김혜진 · 이철희
2	한국경제의 추세 성장을 하락과 원인	석병훈 · 이남강
3	Financial Globalization: Effects on Banks' Information Acquisition and Credit Risk	Christopher Paik
4	The Effects of Monetary Policy on Consumption: Workers vs. Retirees	Myunghyun Kim · Sang-yeon Song
5	북한지역 토지자산 추정에 관한 연구: 프레임워크 개발 및 탐색적 적용	임송
6	김정은 시대 북한의 금융제도 변화 - 북한 문헌 분석을 중심으로 -	김민정 · 문성민
7	Chaebols and Firm Dynamics in Korea	Philippe Aghion · Sergei Guriev · Kangchul Jo
8	한국의 화폐환상에 관한 연구	권오익 · 김규식 · 황인도
9	재원조달 방법을 고려한 재정지출 효과 분석 : 미국의 사례를 중심으로	김소영 · 김용건
10	The Impact of Geopolitical Risk on Stock Returns: Evidence from Inter-Korea Geopolitics	Seungho Jung · Jongmin Lee · Seohyun Lee
11	Real Business Cycles in Emerging Countries: Are Asian Business Cycles Different from Latin American Business Cycles?	Seolwoong Hwang · Soyoung Kim
12	우리 수출의 글로벌 소득탄력성 하락 요인 분석	김경근
13	북한의 경제체제에 관한 연구: 실태와 평가	양문수 · 임송

14	Distribution–Dependent Value of Money: A Coalition–Proof Approach to Monetary Equilibrium	Byoung–Ki Kim · Ohik Kwon · Suk Won Lee
15	A Parametric Estimation of the Policy Stance from the Central Bank Minutes	Dong Jae Jung
16	The Immigrant Wage Gap and Assimilation in Korea	Hyejin Kim · Chulhee Lee
17	Monetary Non–Neutrality in a Multisector Economy: The Role of Risk–Sharing	Jae Won Lee · Seunghyeon Lee
18	International Transmission of Chinese Monetary Policy Shocks to Asian Countries	Yujeong Cho · Soyoung Kim
19	The Impact of Robots on Labor Demand: Evidence from Job Vacancy Data for South Korea	Hyejin Kim
20	전공 불일치가 불황기 대졸 취업자의 임금에 미치는 장기 효과 분석	최영준
21	Upstream Propagation of the U.S.–China Trade War	Minkyu Son
제2022 –1	Immigration and Natives’ Task Specialization: Evidence from Korea	Hyejin Kim · Jongkwan Lee
2	Transmission of Global Financial Shocks: Which Capital Flows Matter?	Bada Han
3	Measuring the Effects of LTV and DTI Limits: A Heterogeneous Panel VAR Approach with Sign Restrictions	Soyoung Kim · Seri Shim
4	A Counterfactual Method for Demographic Changes in Overlapping Generations Models	Byongju Lee
5	Housing Wealth, Labor Supply, and Retirement Behavior: Evidence from Korea	Jongwoo Chung

-
- | | | |
|----|---|---|
| 6 | Demand Shocks vs. Supply Shocks: Which Shocks Matter More in Income and Price Inequality? | Seolwoong Hwang · Kwangwon Lee · Geunhyung Yim |
| 7 | Financial Literacy and Mutual Fund Retail Investing: Evidence from Korea During the 2008 Financial Crisis | Jongwoo Chung · Booyuel Kim |
| 8 | Exchange Rate Regime and Optimal Policy: The Case of China | Yujeong Cho · Yiping Huang · Changhua Yu |
| 9 | 북한 수출입단가지수 추정: 북중무역 데이터를 중심으로 | 이종민 · 김민정 |
| 10 | 탄소배출을 감안한 국가별 녹색 총요소생산성 분석 | 안상기 |
| 11 | 북한 소비자 지급수단 조사 및 분석 | 이주영 |
| 12 | Selection into Outsourcing versus Integration Strategies for Heterogeneous Multinationals | Sangho Shin |
| 13 | Central Bank Digital Currency and Privacy: A Randomized Survey Experiment | Syngjoo Choi · Bongseop Kim · Young Sik Kim · Ohik Kwon |
| 14 | Technological Change, Job Characteristics, and Employment of Elderly Workers: Evidence from Korea | Jongwoo Chung · Chulhee Lee |
| 15 | Machine-Learning-Based News Sentiment Index (NSI) of Korea | Beomseok Seo · Younghwan Lee · Hyungbae Cho |
-

