IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Extracting economic sentiment from news articles: the case of Korea[1]

Younghwan Lee and Beomseok Seo,
Bank of Korea

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Extracting Economic Sentiment from News Articles: The Case of Korea

Younghwan Lee, Beomseok Seo

Abstract

In this study, I propose a News Sentiment Index as an approach to meet the growing need for timely economic statistics. Using a set of machine learning techniques, economic sentiments were extracted from news articles from 2005 to the present to construct this new index. The proposed index complements existing economic statistics in two ways. First, unlike many existing macroeconomic data, it can be available on a daily basis. Owing to recent advances in information technology, this index can be calculated quickly, whereas the calculation of traditional macroeconomic indices is time-consuming and costly. Second, the News Sentiment Index is a good predictor of important macro-variables; not only is it highly correlated with GDP, the Economic Sentiment Index, the Consumer Sentiment Index, and the Business Sentiment Index, but it also leads those indices by one to two months. Empirical evidence supports the hypothesis that news articles convey valuable information regarding economic prospects. The accompanying brief analysis of the 2020 COVID-19 crisis in South Korea proves the usefulness of the index.

Keywords: Big-data analysis, Economic sentiment, Sentiment index, Natural language processing, COVID-19

JEL classification: C45, E37

Table of Contents

1. Introduction

In this paper, a method for constructing an economic sentiment index using news articles is proposed, and the empirical validity of the new index is presented. In contrast to existing survey-based sentiment indexes such as the University of Michigan Consumer Sentiment Index (MCSI) of the US, and the Composite Consumer Sentiment Index (CCSI) of the Bank of Korea, the News Sentiment Index (NSI) uses web-scraped news articles as input data and applies a set of machine-learning techniques to extract economic sentiment from news articles. The two fundamental benefits of substituting survey data with news articles are timeliness and cost efficiency. To use survey data to measure economic sentiment, a large-scale survey must be conducted on a regular basis, which is a time-consuming and expensive task. Furthermore, the survey respondents have to be carefully selected to ensure the representativeness of the analysis and maintained for consistency of the index. On the other hand, a large volume of news articles focused on economic issues is created and stored on a daily basis and is easily accessible via the Internet. Streamlining of the entire process in this way allows us to access economic sentiment quickly with minimal effort.

Despite the apparent advantage of using text data to measure economic sentiment, it is difficult to apply in practice. The difficulties stem from two factors: the unobservability of sentiment and the unstructured nature of text data.

First, the unobservability of economic sentiment raises a question regarding what to measure. The idea of the NSI is to quantify the difference between the number of positive-sentiment sentences and the number of negative-sentiment sentences for a given time period. It is based on the assumption that economic sentiment and the sentiment revealed in economic news articles are highly correlated. Using Michigan Survey data Barsky and Sims (2012) show that consumer confidence reflects the prediction of the change in economic fundamentals. Thus, the assumption is justifiable as long as the economic news articles are based on facts. Specifically, for a given time period t, the proportion between the number of positive sentences and the number of negative sentences is quantified as follows:

$$\mathrm{X_t} = \frac{N\ of\ Positive\ Sentences - N\ of\ Negative\ Sentences}{N\ of\ Positive\ Sentences + N\ of\ Negative\ Sentences}. \quad (1)$$

The NSI at time t is defined as a scaled and translated version of $\mathrm{X_t}$ such that its long-term mean and standard deviation are equal to 100 and 10, respectively.

Second, because the input data are unstructured text data, a question arises as to how to measure the sentiment of each article sentence. A growing body of economic literature is concerned with incorporating text as an alternative data source. Shapiro et al. (2020) provide the closest example to this study. They proposed a method to extract the sentiment of news articles based on sentiment-labeled lexicons and grammatical rules, such as negation, to construct a new sentiment index using news articles. They show that the new index predicts existing sentiment indices such as the MCSI and Conference Board Consumer Confidence. The Economics Policy Uncertainty Index suggested by Baker et al. (2016) is another prominent example. They defined three groups of keywords related to economics, policy, and uncertainty. Their index is based on the proportion of articles that contain a predefined set of keywords.

In contrast to the aforementioned studies, the NSI is based on a machine-learning approach. Rather than using predefined rules, the sentiment classification rule is trained using data and algorithms. The advantage of this approach is its flexibility. The meaning of words is subject to change. For example, in economic news articles, the meaning of word "corona" has changed from a brand of beer to a name of virus that triggered severe economic slowdown. The machine learning approach allows for such changes to be updated via additional training.

A brief analysis provides empirical evidence that news articles convey valuable information regarding economic prospects. The NSI is a good predictor of important macroeconomic variables; not only is it highly correlated with the GDP, CCSI, Business Survey Index (BSI) and Economic Sentiment Index (ESI), but it also leads those indices by one to two months.

The rest of this paper is organized as follows. In Section 2, the methodology for obtaining the sentiment classifier is presented. Section 3 presents a brief analysis of the empirical validity of the NSI. Finally, Section 4 concludes the paper.

2. Methodology

The sentiment classifier is an essential component of the entire process undertaken in this study. Mathematically, it is a function f that takes a sentence as an input and returns the corresponding sentiment label. That is,

$$f : S \rightarrow \{Positive, Negative, Neutral\} \qquad (2)$$

Here, S is the set of article sentences. The question is how to properly approximate the function f. To make it tractable, the classifier function f is deconstructed into two parts: preprocessing and training. The first part, preprocessing, is a step that converts article sentences into sequences of tokens so that they can be used as input for the training step. That is, the preprocessing is a function h such that

$$h : S \rightarrow V^M,$$

where V is the set of tokens, and M is the maximum length of the sequences. All sequences shorter than M are padded with null tokens until their length is equal to M.

The second part, the training, is to find a function that takes a sequence representation of a sentence as input and returns its corresponding sentiment label. Let $g(\cdot|\theta) : V^M \rightarrow \{Positive, Negative, Neutral\}$ be a classifier function parameterized by θ. Assume that it is a true classifier if $\theta = \theta_0$. Equation (2) can then be rewritten as follows:

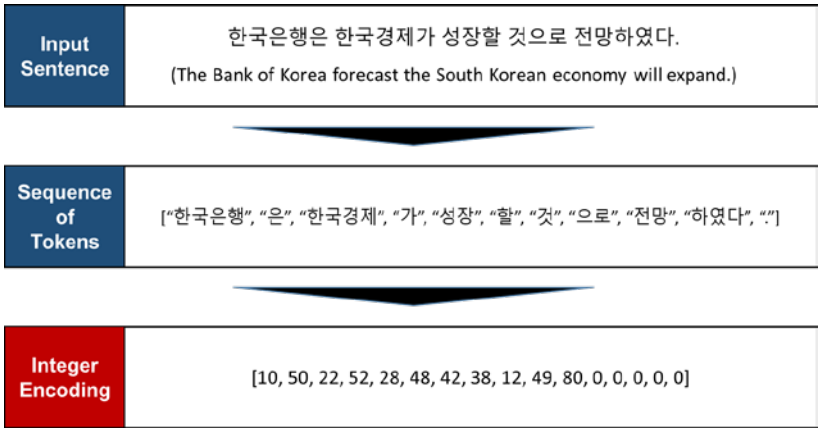
$$f(s_i|\theta_0) = g(h(s_i)|\theta_0),$$

where $s_i \in S$. Thus, given the preprocessor h, the remaining task is to find a proper proxy $\hat{\theta}$ of θ_0 to determine an approximation of f. It is performed by supervised learning with the transformer model proposed by Vaswani et al. (2017) and 450,000 lines of a human operator-labeled training set.

The obtained classifier is applied to categorize the sentiments of the input sentences. Based on the categorization result, the counts of positive and negative sentences are aggregated to calculate and update the NSI. The details are discussed in Section 3.

Example of Preprocessing[1]

Figure 1.



[1] In this example, integer encoding is assumed in which the maximum length of a sequence is equal to 16 and the null token is represented by 0.

Source: Bank of Korea

2.1 Preprocessing

The objective of preprocessing is to convert the article sentences into token sequences. It begins with the tokenization of the target text. Tokenization is the process of separating text into smaller units called tokens, which is an atomic unit of analysis. A token can be a word, morpheme, or character. For example, the sentence "It is a pen" can be tokenized into a list of elements such as ["it", "is", "a", "pen"] or ["i", "t", "i", "s", "a", "p", "e", "n"]. The algorithm and the granularity of tokenization must be determined depending on the goal of the analysis [1]. In this study, tokenization is based on morphemes to account for the nature of the Korean language.

Next, each token is converted into a numeric type, such as an integer value or a one-hot vector. That is, V is not a set of labels of tokens themselves but, rather, a numerical version of it. This step is necessary because the transformer algorithm does not operate directly on label data. Figure 1. summarizes the process.

2.2 Training

The transformer model (Vawani et al., 2017) was used to approximate the second component of the sentiment classifier $g(\cdot|\theta)$. This model has distinctive advantages compared to existing alternatives. First, it improves the accuracy of the support vector machine (SVM) in the sense that it takes positional and contextual information into account. The meaning of a sentence, a sequence of tokens, is more than the sum of the meanings of individual tokens. Rather, the meaning of a sentence is dependent on the order of words as well as on the other words included in the sentence. The transformer model considers positional information and the dependency between tokens to embed input sentences into a semantic space.

[1] For a detailed discussion, see Webster and Kit (1992).

Performance of the Classifier[1]

Table 1.

		Human operator label	
		Positive	Negative
Predicted label	Positive	0.95	0.05
	Negative	0.03	0.97

[1] To evaluate the accuracy of the classifier when it is used to calculate the NSI, the case where either l_i or $\hat{l}_i$ is neutral is excluded from the test.

Source: Bank of Korea

Second, the transformer model improves the speed of the recurrent neural network (RNN) model because it allows for more parallelization in the computation process. The RNN model sequentially takes each token in a sequence and embeds it into a semantic space. The sequential nature of the model allows it to preserve the positional information of a sentence at the cost of computation time. However, the transformer model does not sequentially process the input sequence to incorporate the positional information of tests. Rather, it relies on an attention mechanism that can be easily parallelized. This characteristic enables the transformer model to exploit parallel computing power to speed up training.

Approximately 450,000 labeled instances were used to train the transformer model. Each labeled instance is the sentence-sentiment label pair $(s_i, l_i) \in S \times \{Positive, Negative, Neutral\}$ where each sentiment label l_i is manually labeled by human operators. The training sentences were randomly selected from economic news articles from 2005 to 2021. Based on the training set and the transformer model, the proxy $\hat{\theta}$ of θ_0 is calibrated. The composition of $g(\cdot)$ and $h(\cdot\,|\hat{\theta})$ results in the classifier $\hat{f}(\cdot)$.

To test the validity of the classifier, an additional 5,000 sentences were randomly sampled and labeled by human operators to construct the validation set. Table 1 summarizes the performance of the classifiers. Each entry is the corresponding conditional probability $\Pr[\,\hat{l}_i \mid l_i]$ where $\hat{l}_i = \hat{f}(s_i)$. Overall, it showed an acceptable level of reliability.

3. Empirical Result

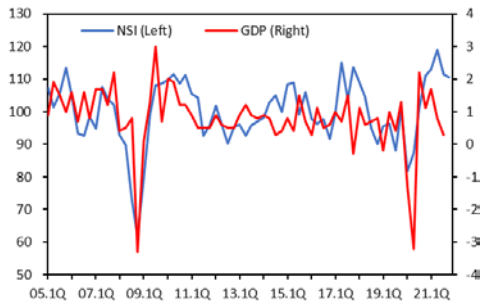
To construct the target dataset for the NSI, all news articles accessible on the economic section of a web portal were crawled on a daily basis. It covers 3,500 articles from 50 newspapers a day on average as of 2021. After separating articles into sentences, 10,000 were randomly selected as the target sample. The time coverage of the dataset is from 2005 to the present, which includes several major events such as the financial crisis and COVID-19.

The classifier was applied to categorize the target sentences. For each day, the number of positive and negative sentences was counted. Using this sentiment count, the NSI can be calculated for an arbitrary period. For example, by applying equation (1) to the monthly and quarterly sum, the monthly and quarterly NSI is calculated after standardization. On the other hand, the daily sentiment NSI is based on the sum of the past seven days to remove the day of the week effect.

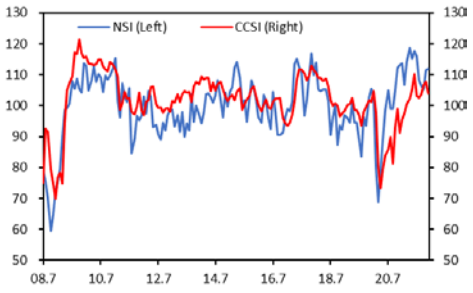
Macro-Variables with the NSI

Figure 2.

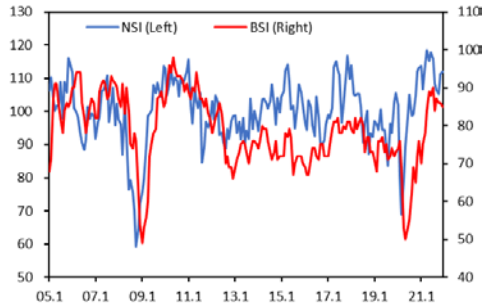
GDP[1] with the Quarterly NSI



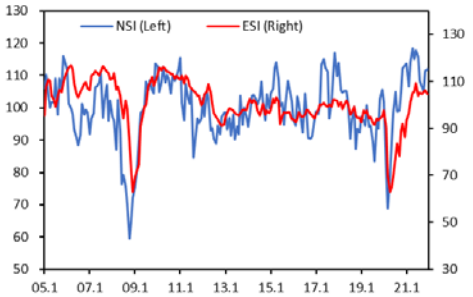
CCSI with the Monthly NSI



BSI with the Monthly NSI



ESI with the Monthly NSI



[1] GDP here is the seasonally adjusted real GDP growth rate.

Source: Bank of Korea

To examine the empirical validity of the NSI, the correlations between the NSI and four macro-variables were analyzed: Gross Domestic Product (GDP), Composite Consumer Sentiment Index (CCSI), Future Business Condition Business Survey Index (BSI), and Economic Sentiment Index (ESI). Figure 2 shows a plot of the NSI and the macro-variables. The NSI closely follows all of these variables. The findings are twofold. First, GDP and the quarterly NSI are highly correlated. The correlation coefficient of the two variables was 0.55. This implies that news articles convey substantial information regarding economic fundamentals.

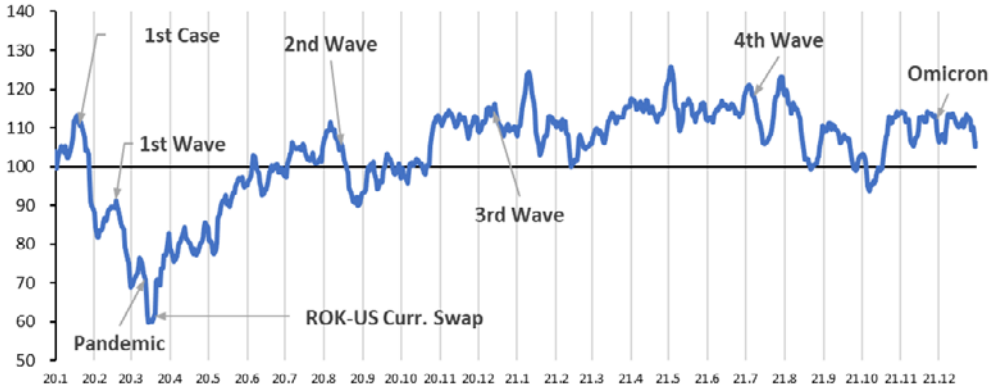
Second, the lag correlation analysis between the monthly NSI and sentiment-related macro-variables[2] suggests that the NSI can help econometricians predict changes in economic sentiment. The results of the analysis show that the NSI leads CCSI by one month, BSI by two months, and ESI one month, where the corresponding maximum correlation coefficients are 0.75, 0.61, and 0.61, respectively.

In Figure 3, the daily NSI and the dates of several important COVID-19-related events are indicated. This shows that the timeliness of the index helps identify the impact of specific events on economic sentiment. For example, after the first confirmed case was reported in the Republic of Korea, the NSI dramatically decreased until the Bank of Korea (BOK) and the US Federal Reserve signed a currency swap contract. After the contract, the NSI improved for five months until the second wave came.

[2] CCSI, BSI, and ESI.

Covid-19 with the daily NSI[1]

Figure 3.



[1] The daily NSI here is constructed based on the news articles of the previous seven days.

Source: Bank of Korea

4. Concluding Remarks

In this paper, a methodology to construct a sentiment index based on internet news data was proposed, and it was found that the new index (NSI) predicts the existing macro-variables, including sentiment indices. The new index allows policymakers to access current economic sentiment in a timely manner with little marginal effort because it is not based on a survey, which is costly to conduct.

Understanding the topics that drive economic sentiment remains a topic for future study. Refining the keywords exposed in the positively or negatively classified sentences will allow us to keep up with current events that drive current economic sentiment.

References

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131* (4), 1593-1636.

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

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

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998-6008

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

Short Bio

Younghwan Lee is an economist with the Bank of Korea, which joined in September 2019. He previously worked as an assistant research professor at Seoul National University. He earned his PhD in economics from Seoul National University and an undergraduate degree in Management Science from the Korea Advanced Institute of Science and Technology. His research interests lie in the fields of computational economics and finance.

Beomseok Seo is an economist in the Bank of Korea, where he joined in January 2011. He earned his PhD in statistics from Pennsylvania State University researching interpretable machine learning and bachelor's degree in economics and statistics from Korea University. His research interests lie primarily in the fields of statistical modeling and machine learning for the better human interpretation.

Extracting Economic Sentiment from News Articles: The Case of Korea

Younghwan Lee, Beomseok Seo

Economic Statistics Department,
Bank of Korea

February 10, 2022

Motivation

- Currently, economic conditions change very quickly.
- A timely assessment of economic conditions is especially valuable when the market is highly volatile and uncertain.
- Does official statistics fast enough? Think about survey based statistics such as Michigan Consumer Sentiment Index (MCSI).
- Collecting relevant data is time consuming and costly.
- On the other hand, tons of data is produced and stored at every second. Internet news articles data is one of the examples.
- This study proposes the ***News Sentiment Index (NSI)*** which is a timely available and cost-efficient measurement of economic sentiment.

The Idea

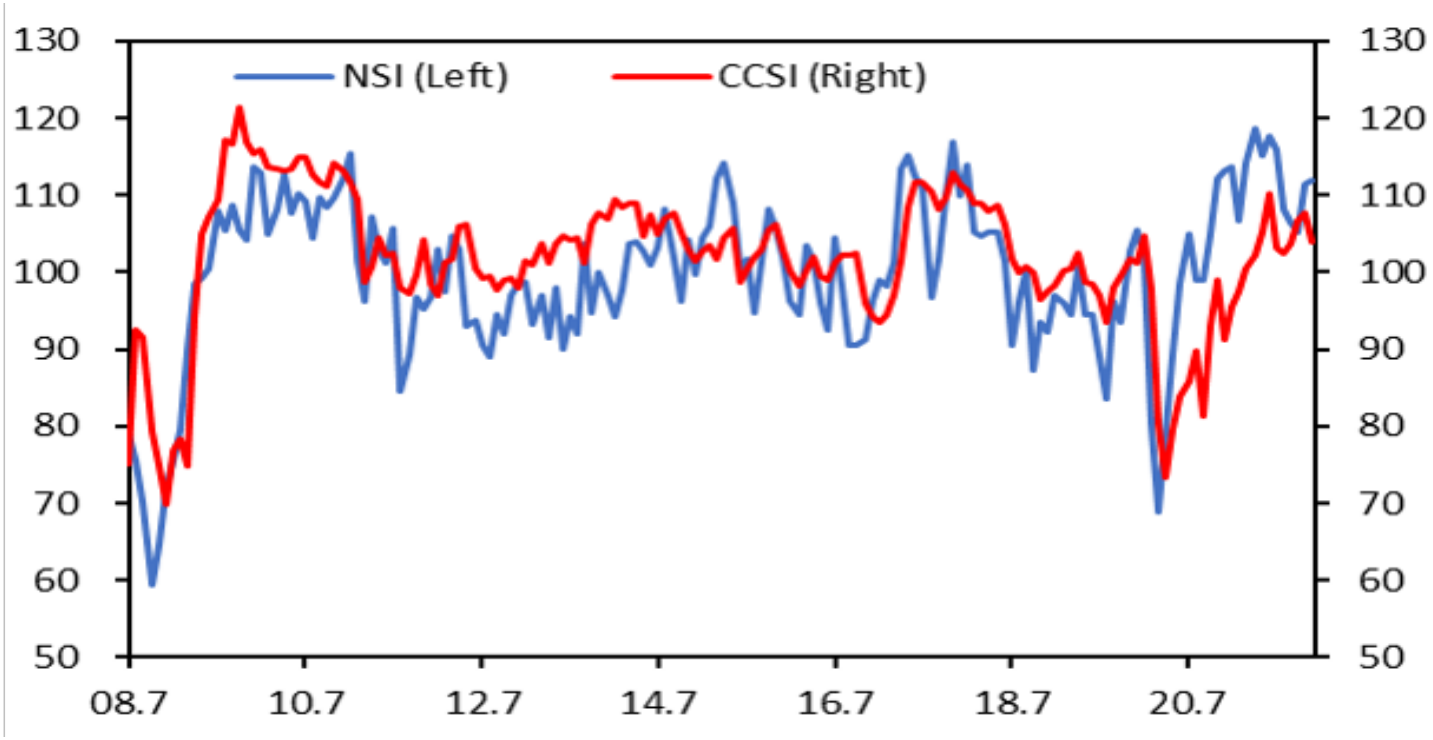
- Randomly sample 10,000 article sentences on a daily basis and classify their sentiments into three categories: positive, negative, and neutral.
- Counting period can be arbitrarily chosen.
- Quantify the number of positive and negative sentences of news articles as follow:

$$X_t = \frac{\#\,\text{of pos. sentences} - \#\,\text{of neg. sentences}}{\#\,\text{of pos. sentences} + \#\,\text{of neg. sentences}}$$

- Translate and scale X_t to make it looks like an index. (Mean = 100, Std. = 10)

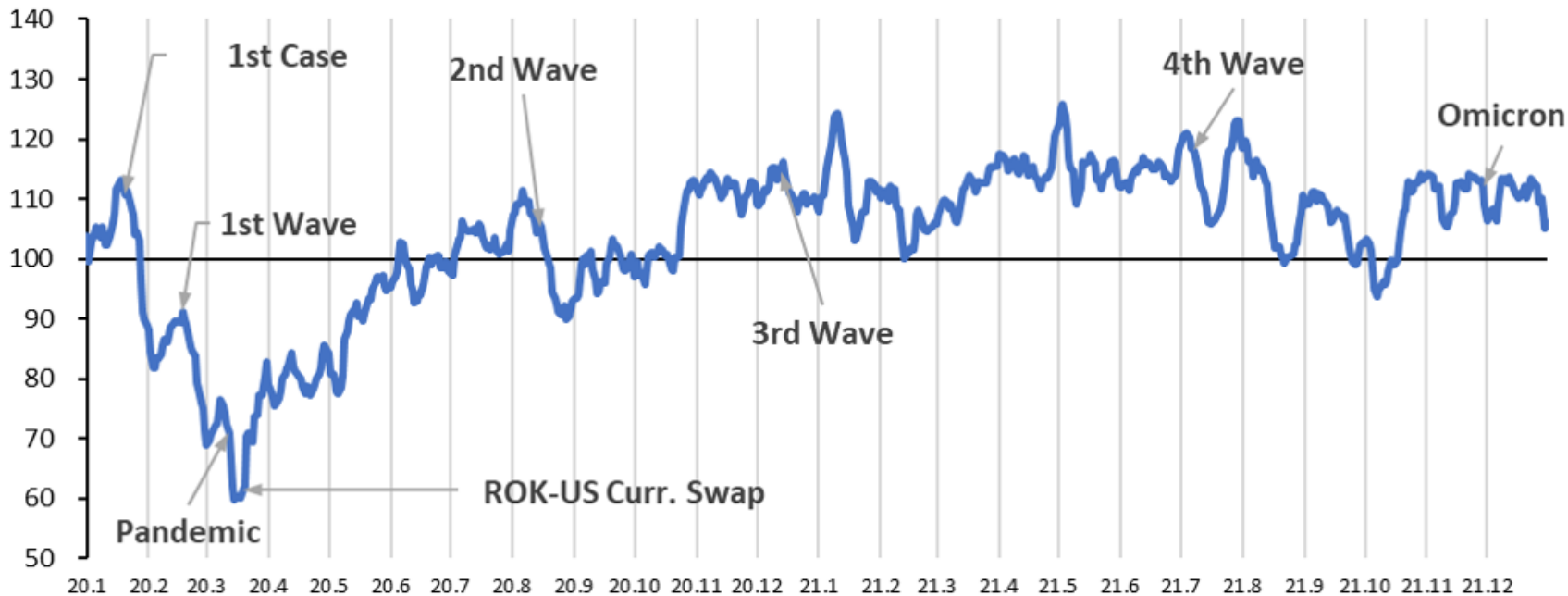
CCSI and NSI

- Composite Consumer Sentiment Index (CCSI) is a survey based consumer sentiment index in Korea.
- The monthly NSI lead it by 1 month and correlation is 0.75.



COVID-19 and NSI

- The daily NSI is able to react quickly to changes in economic conditions such as COVID-19 events.



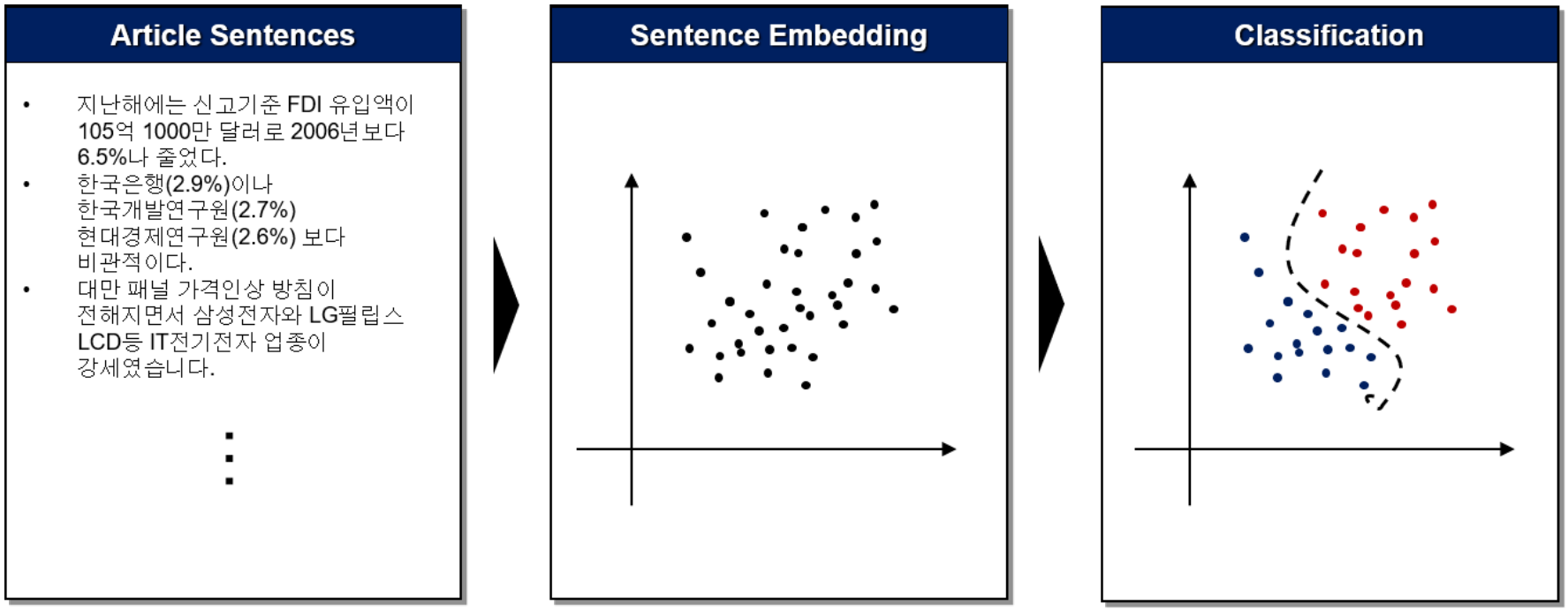
Methodology: Overview

- Supervised learning approach:
  - Input: 450,000 human operator labelled sentence-sentiment pairs.
  - Model: Transformer model
- Different from lexical approach from Shapiro et al.(2020), this approach is:
  i) does not require pre-defined rules.
  ii) change in meaning can be updated by additional training. (Ex. Corona vs Corona)

- Work flow:

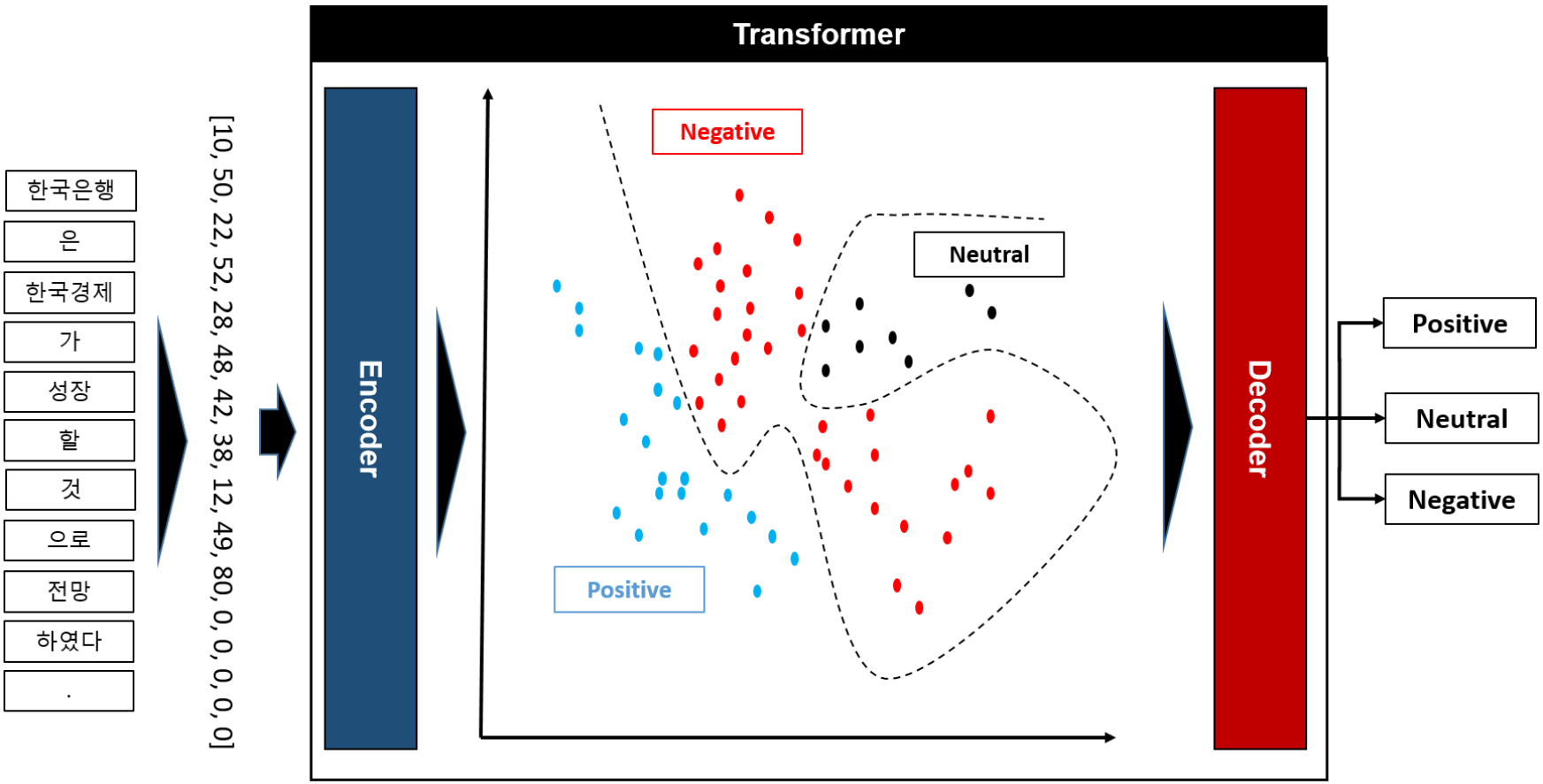
Methodology: Intuition

- We cannot directly apply classification models to the text data.
- The idea is to embed sentences into the ***semantic space*** that preserves senses of sentences.



Methodology: Transformer Model

- Vaswani et al. (2017) have proposed the transformer model.
- It takes positional information into account while faster than RNN.



Concluding Remarks

Implication:

- By using big-data, we can complement existing official statistics in terms of cost and time.
- Unstructured data such as text can help us to make better policy decision

Future work:

- What is the driving force of the change in economic sentiment?