Extracting Economic Sentiment From News Articles: The Case of Korea

Younghwan Lee

Bank of Korea

September 21, 2023

Disclaimer: The views expressed in this talk do not reflect necessarily those of Bank of Korea

Motivation

- To produce official statistics, the collection of relevant data is a time-consuming and costly endeavor. Consider survey-based statistics, such as the Michigan Consumer Sentiment Index (MCSI).
- On the contrary, vast amounts of data are constantly generated and stored every second, including internet news articles.
- In this presentation, I will introduce methodologies for extracting information from news articles with application example of Korea. Specifically, it includes followings:
  i) The ***News Sentiment Index (NSI)*** which is a timely available and cost-efficient measurement of economic sentiment based on news articles.
  ii) The ***Trending Keywords*** analysis to closely monitor up-to-date economic issues.

News Sentiment Index: Idea

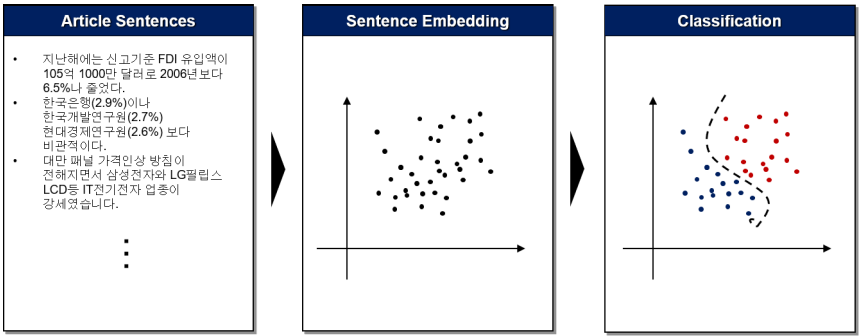
- Randomly sample 10,000 article sentences on a daily basis and classify their sentiments into three categories: positive, negative, and neutral.
- Counting period can be arbitrarily chosen.
- Quantify the number of positive and negative sentences of news articles as follow:

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

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

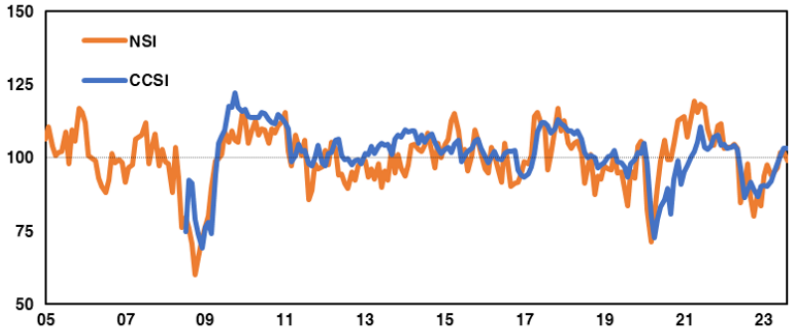
News Sentiment Index: Methodology

- 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.



News Sentiment Index: Result

- Monthly NSI lead Composite Consumer Sentiment Index (CCSI) by 1 month and their correlation coefficient is 0.75.



* NSI was recognized by KOSTAT as an experimental statistic in February 2022.

Trending Keywords: Idea

- ***Trending Keywords*** are defined as keywords extracted from news headlines during a specific time period that best describe the key issues of that period.
- Then, how can we capture the ***Trending Keywords***?
- Suppose that you are provided with a collection of news headlines, and you are asked to guess when those headlines were published based on the keywords in the headlines.
- I define ***Trending Keywords*** as a set of keywords that are most useful for us to address the question above.

Trending Keywords: Methodology

- The Neyman-Pearson lemma states that the log-likelihood test is the most powerful test. By utilizing the empirical likelihood ratio of keyword frequency, we can address the question at hand.

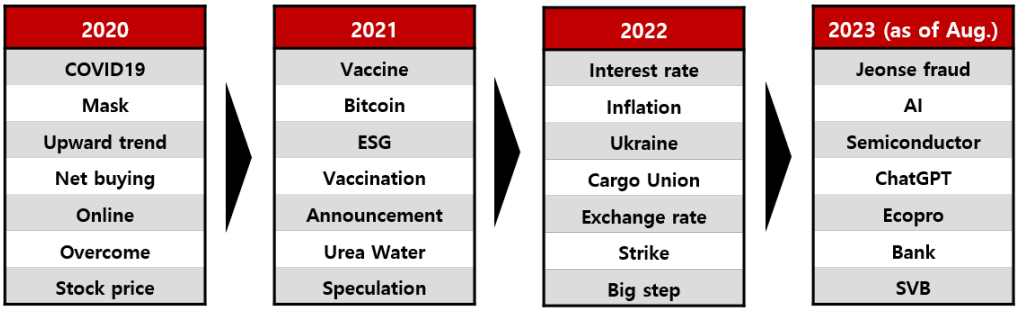
- For each keyword $i \in I$ define trending score S_i as follow:

$$S_i = \Pr[i|\text{target period}] \ln \frac{\Pr[i|\text{target period}]}{\Pr[i|\text{base period}]}.$$

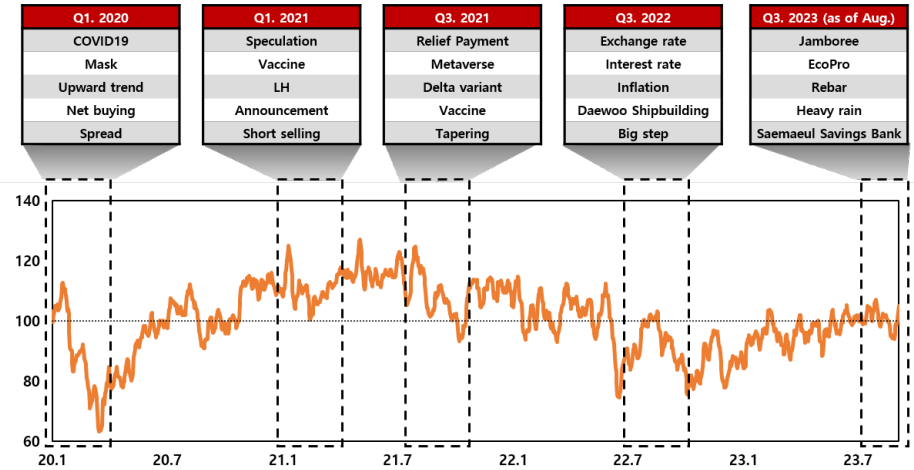
- S_i increase when the keyword i
  - i) is encountered more frequently in the headlines during the target period,
  - ii) is relatively uncommon in the base period.
- We can extract ***Trending Keywords*** for the target period by ranking keywords according to their trending score.

Trending Keywords: Result

- The yearly ***Trending Keywords*** from 2020 summarize the economic consequences of the COVID-19 outbreak: asset market bubble, supply bottleneck, and turmoil in the banking sector.



NSI with Trending Keywords



Concluding Remarks

- The results of this study suggest that news articles can provide valuable information for monitoring the economic condition.
  - ***NSI*** conveys information about the direction of economic sentiment.
  - ***Trending Keywords*** summarize the up-to-date economic issues.

- Future research:
  - Is it possible for the NSI to be domain-specific? (ex. specific region)
  - What are the driving forces behind keyword generation?
  - Can we predict future keywords?