News-based Economic Report Generation for Business Sentiment Index

Kazuhiro Seki
Faculty of Intelligence and Informatics
8-9-1 Okamoto, Higashinada, Kobe
Hyogo, Japan
e-mail: seki@konan-u.ac.jp

***Abstract*—Business sentiment indices (BSIs) derived from surveys play a crucial role in economic decision-making but suffer from high costs and limited timeliness. While recent work has shown that BSIs can be estimated from news articles more efficiently, such indices lack interpretability, making it difficult to understand the economic factors behind their values. This paper investigates methods for generating concise and interpretable monthly economic reports to explain news-based BSIs by leveraging large language models (LLMs). We compare three methods: direct one-shot report generation, one-shot report generation with topic categorization via prompting, and a topic-sentiment-driven method that integrates neural topic modeling, sentiment aggregation, and multi-stage LLM-based report generation. Experiments on one year of Japanese web news demonstrate that the topic-aware prompting method produces the most distinctive topics but yields less accurate economic impact scores than the topic-modeling method, which reduces API cost but is slower, highlighting trade-offs among report quality, topic diversity, API cost, and processing time.**

Keywords-text summarization; large language models; sentiment analysis; topic analysis.

I. Introduction

Recent advances in natural language processing have enabled the automated generation of economic reports to assist in interpreting business sentiment indices (BSIs) [1]. Business sentiment indices (BSIs) serve as key references for a wide range of economic activities, including monetary-policy decisions, corporate production planning, and investment decisions by institutional and individual investors. In Japan, widely used BSIs include the Current and Future Assessment Diffusion Index of the Economy Watchers Survey [2] by the Japanese government and the Diffusion Index of the Tankan Survey [3] by the Bank of Japan. However, the surveys underlying the computation of these indices require substantial time, labor, and financial resources.

For example, in the case of the Current and Future Assessment Diffusion Index of the Economy Watchers Survey (hereafter, the Economy Watchers DI), regional survey organizations administer questionnaires to 2,050 individuals, such as retail store owners and taxi drivers, who are well positioned to observe local economic conditions. These surveys cover 12 regions of Japan, and the collected regional results are aggregated and analyzed by a central coordinating organization to compute the final index values. Because such surveys take considerable time to conduct and process, the Economy Watchers DI is released only once a month and the Tankan Survey DI only once a quarter, which limits their timeliness.

To address these limitations, Seki et al. [4] developed a framework to estimate a BSI from web news articles automatically collected from the Web. While this framework showed a strong correlation with a traditional survey-based index—specifically, the Economy Watchers DI—and offered significant improvements in timeliness and cost-efficiency, it lacked interpretability. That is, it is unclear why a particular BSI value was obtained.

To bridge this gap, the present study extends the preliminary work by Seki [1]. We investigate methods that leverage sentiment and topic analyses to capture sentiment distribution and topic structure, followed by text summarization to synthesize relevant economic events from news articles. Experiments on news articles collected in 2025 demonstrate promising results, producing concise and easily interpretable summaries of the economic situation for each period.

The rest of this paper is structured as follows: Section II reviews related work. Section III describes our proposed approach, Section IV presents experimental results on real-world news data, Section V introduces the functionality to display economic reports with our BSI nowcasting system, and Section VI concludes with a summary and possible future work.

II. Related Work

This section summarizes related work on forecasting economic indices from textual data and on economic report generation.

There have been a number of attempts to forecast or nowcast economic indices, including BSIs, from textual data [4]–[6]. For example, Seki et al. [4] nowcast BSI using daily Japanese newspaper articles. They used newspaper sentences as input to a regression model to estimate the economic impact score of each sentence. The regression model was a BERT model fine-tuned on the Economy Watchers Survey data in Japan. The economic impact scores were averaged monthly to obtain the BSI for each month. The resulting BSI values showed a strong correlation ($R = 0.937$) with an existing survey-based BSI. Seki [7] also examined the influence of newspaper-specific biases on BSI computation and applied a dynamic factor model [8] to obtain an unbiased BSI using multiple newspapers.

In other languages and countries, Bieri et al. [9] constructed a news-based sentiment indicator from multilingual newspapers (German- and French-language sources) and examined its relationship with existing survey-based indicators. Kalamara et al. [10] demonstrated improvements in forecasting GDP and related macroeconomic variables using UK newspaper texts, while Ashwin et al. [11] showed that sentiment derived from newspaper text can contribute to improved nowcasting of euro area GDP. Furthermore, Barbaglia et al. [12] conducted aspect-based, fine-grained sentiment extraction and reported enhanced predictive performance for multiple macroeconomic variables.

Generating an economic report from multiple news articles can be seen as multi-document summarization (MDS), a topic with extensive literature; thus, the following focuses on economic-report generation and then touches upon topic-guided MDS. Hu et al. [13] proposed a multi-edit Bi-LSTM neural network that first learns an outline for the input news and then generates financial reports from the learned outline. Yan et al. [14] focused on topic accuracy and proposed a variational topic inference model that improves topic faithfulness in report generation by incorporating latent variables. Zhang et al. [15] also proposed a method that improves faithfulness in multi-document summarization by integrating relationships among entities and events extracted from multiple documents into a cross-document information extraction graph and incorporating this structural information into the summarization process. Furthermore, given recent advances in large language models (LLMs), there is growing interest in using LLMs for report generation. For instance, Shukla et al. [16] first identified crucial narrative sections within input financial reports and then provided the extracted sections to LLMs for summarization. Yang et al. [17] conducted a comparative study of three LLMs (GLM-4, Mistral-NeMo, and LLaMA3.1), focusing on their effectiveness in generating automated financial reports. Assis et al. [18] used LLMs to automatically generate market commentary from corporate disclosure documents for the Brazilian financial market.

For MDS, the use of topic modeling has also been widely investigated. Cui and Hu [19] proposed an MDS framework that constructs a heterogeneous graph consisting of word, topic, and document nodes, and integrates latent topics estimated by a neural topic model into summary generation as global semantics that bridge across documents. Their method jointly learns summarization and topic estimation, and introduces an inconsistency loss to encourage alignment between the topic distributions of the input documents and the generated summaries. Dong et al. [20] formulated MDS as a two-stage process: subtopic detection based on content attention in the first stage, followed by a submodular optimization problem in the second stage that simultaneously minimizes inter-sentence distance and maximizes topic diversity. Li et al. [21] proposed topic-guided reinforcement learning, which extracts topic phrases or labels to construct a topic-based reward for preserving topic relevance in MDS, and optimizes the summarization model using the Group Relative Policy Optimization (GRPO) framework. The reward defines topic alignment in terms of both coverage and precision.

The present work differs from the topic-centered text generation studies summarized above in that it aims to generate an economic report guided not only by topics but also by their economic impacts.

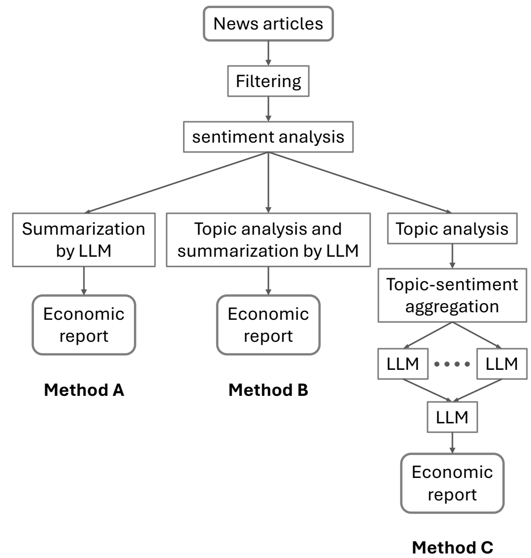

Figure 1. Overview of the three methods.

III. Proposed Approach

A. Overview

We devised three methods to address the limited interpretability of a news-based BSI by generating an economic report from the news articles used to compute the BSI. Figure 1 depicts the processing flows of the three methods. The first method (Method A) generates a report directly from the input news articles using a large language model (LLM). The second method (Method B) also uses an LLM, but first categorizes news articles into representative topics and then summarizes a limited number of major topics using a single detailed prompt. The third method (Method C) employs a separate topic analysis model and aggregates the outputs of topic and sentiment analyses; these outputs are then provided to LLMs to generate topic-wise summaries, which are finally combined by another LLM to produce an economic report.

In the following, we first summarize how BSI is computed, then describe the input data for the three methods, and lastly detail each method.

B. BSI Computation

For BSI computation, we basically follow the framework developed by Seki et al. [4]. The primary difference is that Seki et al. used a fine-tuned BERT model, whereas we adopt ModernBERT [22] as a base model. For completeness, the procedure to compute BSI is summarized as follows:

1) Automatically collect Japanese web news articles from the Sankei Newspaper website [23].
2) Split each news article into sentences based on Japanese periods.
3) Filter out irrelevant news sentences by applying an outlier detection model based on a one-class SVM [24] trained on Economy Watchers Survey data.

```
The following are headlines of multiple news articles arranged in
chronological order. Based on sentiment analysis, these texts have
been identified as having a significant impact on economic
conditions. Taking a comprehensive view of these texts, generate an
economic report from the perspective of their impact on the economy.

### Output Requirements
- The output must be in JSON format only.
- Do not include any additional explanatory text, extra keys, or
Markdown tags.
- The value of "overall_summary" must be a string written in
Japanese, starting with "Recent events:" and using the formal
declarative style.
- The length of "overall_summary" should be approximately up to 300
Japanese characters.

### Output Format
{
  "overall_summary": "Recent events: ..."
}

### News Texts
{N} lines of news headlines with sentiment scores
...
```

Figure 2. English-translated prompt used for Method A.

4) Apply a sentiment analysis model to each news sentence to predict its business sentiment score. The model is a Japanese ModernBERT model fine-tuned on the Economy Watchers Survey data. If the model output is positive, the economic impact is considered positive, and vice versa.
5) Compute the monthly average of the sentiment scores as the BSI of the month.

C. Input Data

As input for generating an economic report, it would be straightforward to use the news sentences and their sentiment scores computed as in Section III-B. However, our preliminary experiment showed that news sentences were often too fragmented, lacking sufficient context to be analyzed correctly for economic report generation.

To mitigate the problem, we use news articles without splitting them into sentences and compute their sentiment scores. However, as one month of news articles is too long to be fed to an LLM, we use only the headlines with the sentiment scores computed for the corresponding articles as input to the three methods described below.

D. Method A: One-shot economic report generation

Method A relies entirely on an LLM prompt shown in Figure 2 to generate an economic report. The LLM is instructed to generate a report based on the N input news articles (headlines) and their sentiment scores. The N input news headlines are selected based on the absolute sentiment scores of the news articles computed in Section III-B. It should be mentioned that, unlike the other methods, this method does not provide topic information to be presented in our system (see Section V) and is tested only for comparison purposes.

E. Method B: One-shot economic report generation with topic categorization

Method B takes the same input news headlines as Method A but instructs an LLM to categorize input news under k topics and then to generate an economic report based on the most important $k'(< k)$ topics. The prompt is shown in Figure 3.

```
The following are headlines of multiple news articles arranged in
chronological order. Based on sentiment analysis, these texts have
been identified as having a significant impact on economic
conditions.

### Task
You must execute the following steps in the specified order.

1. Internally classify these texts into {k} topics based on their
content (note: these {k} topics themselves must not be output).
2. For each topic, calculate an "impact_score" by simply summing all
numerical values appearing at the end of the sentences belonging to
that topic (negative values indicate negative impact). The sign must
be preserved.
3. Select the top {k'} topics in descending order of the absolute
value of their impact scores.
4. For each selected topic, generate a summary from the perspective
of its impact on economic conditions.
5. Taking an overview of these {k'} topics, generate a single
integrated economic report that reflects their commonalities and
interrelationships. Avoid a mere enumeration of individual topics;
produce a summary that conveys the overall picture.

### Output Requirements
- The output must be in JSON format only.
- Do not include any additional explanatory text, extra keys, or
Markdown tags.
- The value of "overall_summary" must be a string written in
Japanese, starting with "Recent events:" and using a formal
declarative style.
- The length of "overall_summary" should be approximately up to 300
Japanese characters.
- "topics" must be an array with exactly {k'} elements.
- Each topic must be an object with the following keys:
  - "topic_id": a sequential integer starting from 1, ordered by
descending absolute value of impact_score
  - "label": a concise label of approximately 10–15 Japanese
characters representing the topic
  - "summary": a summary of the topic of up to approximately 300
Japanese characters
  - "impact_score": the calculated numerical impact score

### Output Format
{
  "overall_summary": "Recent events:...",
  "topics": [
    {
      "topic_id": 1,
      "label": "...",
      "summary": "...",
      "impact_score": ...
    },
    ...
    {
      "topic_id": {k'},
      "label": "...",
      "summary": "...",
      "impact_score": ...
    }
  ]
}

### News Texts
{N} lines of news headlines with sentiment scores
...
```

Figure 3. English-translated prompt used for Method B.

The intention behind Method B is to mimic the pipeline of Method C, described next, without an external topic model by using a detailed one-shot prompt to see how well it works compared with Method C.

F. Method C: Topic-sentiment-driven economic report generation

Method C integrates topic modeling and sentiment analysis to generate concise, topic-focused economic report using an LLM. The overall process consists of three main stages: topic modeling, topic-level sentiment aggregation, and LLM-based report generation. In the following topic-modeling context, we treat each news headline as a *document*.

First, all documents are tokenized using a custom Japanese tokenizer, which conducts morphological analysis and extracts compound nouns. The tokenized documents are then used as input to a neural topic model. We adopt FASTopic [25], a fast

```
The following sentences are news headlines arranged in chronological
order and have been classified into the same topic as a result of
topic analysis. The numerical value at the end of each sentence
indicates the strength of its impact on economic conditions
(negative values indicate a negative impact).

Your task consists of the following two items:
1. Generate a single concise topic label of approximately 10-15
Japanese characters that succinctly represents the overall news.
2. From the perspective of the impact on economic conditions,
produce a summary of the entire news in Japanese, written in a
formal declarative style, with a maximum length of approximately 300
Japanese characters.

Output the result only in the following JSON format, and do not
include any other Markdown tags, explanatory text, preambles, or
additional content.

Output Format:
{
  "label": "topic label",
  "summary": "summary focusing on the impact on economic conditions"
}

### News Texts
{N} lines of news headlines with topic-sentiment scores
...
```

Figure 4. English-translated first prompt to generate a topic summary.

```
The following is a report summarizing recent events that are
considered to have had a strong impact on economic conditions for
each of {k'} topics. Each topic is assigned a label, a summary, a
topic ID, and an impact score (negative values indicate a negative
impact).

### Task
Taking an overview of all the topic-specific reports below, generate
a single economic report that reflects their commonalities and
interrelationships. Do not simply list individual topics; instead,
produce an integrated summary that conveys the overall picture.

### Output Requirements
- The output must be in JSON format only.
- The only key allowed is "summary".
- Do not include any additional explanatory text, line breaks,
comments, extra keys, or Markdown tags.
- The value of "summary" must be a string written in Japanese,
starting with "Recent events:".
- The length of the summary should be approximately up to 300
Japanese characters.

### Output Format
{
  "summary": "Recent events: ..."
}

### Topic-wise Reports (JSON format)
{k'} topic-wise reports follows.
...
```

Figure 5. English-translated second prompt to generate an overall summary.

and stable topic model that reconstructs semantic relations from pretrained embeddings using optimal-transport regularization. FASTopic is reported to achieve higher performance in topic coherence [26] and topic diversity [27] than other models, such as LDA [28] and BERTopic [29]. Given a predefined number of topics k, the model is trained on the input documents using dense document embeddings. As output, the model produces (i) a topic-word distribution $p(\mathrm{word}|\mathrm{topic})$ for each topic and (ii) a document-topic distribution $p(\mathrm{topic}|\mathrm{document})$ representing the degree to which each document is associated with each topic.

Second, document-level sentiment scores are combined with the document-topic distribution to compute topic-level sentiment contributions. Specifically, for each document, its topic membership probabilities $p(t_j|d_i)$ are multiplied by its sentiment score s_i, yielding a topic-sentiment matrix, where i and j denote the indices of a document d and a topic t, respectively. By averaging these values over all documents $d_i \in D$, an aggregated topic-sentiment score $a_j = \frac{1}{|D|}\sum_i s_i \times p(t_j|d_i)$ is obtained for topic t_j, which indicates the overall economic impact of the topic. Topics are then ranked by the absolute magnitude of their aggregated topic-sentiment scores $A = (a_1, \ldots, a_k)$, allowing us to identify topics with the strongest positive or negative economic implications.

For each of the k' highly ranked topics among the k topics, documents are further sorted by the absolute value of their topic-sentiment contribution $s_i \times p(t_j|d_i)$. From this ordering, the top n documents are selected as the most influential textual evidence for that topic t_j, and their timestamps are retained to preserve temporal structure.

Next, an LLM is used to generate summaries for each of the k' economically influential topics. The n documents selected for each topic are arranged in chronological order and provided to the LLM as a single prompt shown in Figure 4.

Each document is accompanied by its topic-sentiment score $s_i \times p(t_j|d_i)$, indicating the strength and direction of its economic impact. The LLM is instructed to (i) generate a concise topic label and (ii) produce a short summary that focuses explicitly on the topic's economic implications.

Finally, the k' outputs of the LLM from the previous step are given to an LLM with a prompt to generate an overall economic report covering the major topics. Figure 5 shows the prompt.

IV. Evaluation

A. Experimental Settings

For development and evaluation of the proposed approach, we crawled Japanese news data from the Sankei Newspaper website from October 2024 to December 2025. Note that the full text of the articles is available only to paid subscribers, and our crawled data include only the headlines and the beginning of each article. The length of the crawled text varies across articles, but it is generally sufficient to understand their contents. The 2024 data (October to December) were used for development, and the 2025 data (January to December) were used for evaluation.

For the input data size, we set the number of news headlines (N) to 1,000, considering the LLM context size for Methods A and B and the burden of the evaluator. Note that the LLM context size is not a big issue for Method C since the first stage (i.e., topic analysis) identifies only relevant news articles for each topic but uses the same N for a fair comparison.

As the FASTopic embedding model for Method C, we used "paraphrase-multilingual-mpnet-base-v2". We set the number of topics for FASTopic (k) to 10, the number of topics presented in the final report (k') to 3, and the number of documents per topic (n) to 100. These parameters were set based on a preliminary experiment on the development data, which were also used to refine the prompts for all three methods. Note that k' can be chosen dynamically by setting a threshold on aggregated topic-sentiment scores A. In this study, however, we fixed $k' = 3$ to directly compare the resulting topics identified by Methods B and C.

As an LLM, we adopted OpenAI's gpt4o-2024-08-06 and set the temperature and nucleus sampling parameters to 0 and 1, respectively, for better reproducibility. We also tested a newer model, gpt-5.2-2025-12-11, which sometimes generated awkward Japanese; therefore, it was not used in this experiment.

All experiments were conducted on a MacBook Pro equipped with an Apple M3 Pro chip (11-core CPU), 18 GB of unified memory, with GPU acceleration enabled via Apple's Metal Performance Shaders (MPS).

B. Evaluation Criteria

We compared the reports generated by the three methods in the following aspects:

- Manual evaluation of the k' identified topics. These aspects were evaluated for only Methods B and C as Method A was not designed to select k' topics.
  - *Economic Significance*: whether the selected topics represent the major economic issues of the month, in terms of including economically significant events, not omitting key topics, and avoiding the selection of minor issues as primary topics.
  - *Distinctiveness*: whether the selected topics are sufficiently distinct from one another, in terms of avoiding substantial content overlap and ensuring that each topic represents a clearly different issue.
  - *Impact Score Validity*: whether the assigned economic impact scores for the three topics are reasonable, in terms of the plausibility of their relative magnitudes within the same month based on the topic descriptions.
- Manual evaluation of the generated economic report.
  - *Factual Accuracy*: whether the generated economic report is factually accurate, in terms of not introducing information unsupported by the source news, correctly describing numerical values and events, and avoiding misleading or incorrect statements.
  - *Economic Adequacy*: whether the generated content is adequate as an economic report, in terms of factual consistency with the input topics and source news and the economic plausibility of the interpretation.
- Automatic evaluation of the k' identified topics. This aspect was not evaluated for Method A for the same reason as above.
  - *Diversity*: whether the k' selected topics are diverse, defined as:

$$Diversity = \frac{1}{{}_nC_2} \sum_{i<j} (1 - \cos(t_i, t_j)) \tag{1}$$

where t_i and t_j denote the embedding vectors of the topic summaries for topics i and j ($i \leq k'$, $j \leq k'$), and $\cos(\cdot)$ returns the cosine similarity between the two vectors. The embedding vectors were obtained via the OpenAI API using the text-embedding-3-large model.

- API cost in U.S. dollars, based on the number of input and output tokens. At the time of writing, the cost of gpt-4o was \$2.5 and \$10 per 1 million input and output tokens, respectively.
- Processing time in seconds, measured for the OpenAI API calls and, for Method C only, the topic modeling stage.

The five metrics in the manual evaluation were assessed by an expert with a Ph.D. in Economics on a five-point Likert scale, where higher scores indicate better quality. Before assessing the generated contents, the evaluator first read (1) Jiji Press, Ltd.'s website at https://www.jijiphoto.jp/resource/html/ext/news/year/2025/index.html to grasp the major events that occurred in each month and (2) the actual N news headlines provided to the three methods. More detailed instructions are found in Appendix A.

C. Example Outputs

To illustrate how each method worked and how it was assessed, we first present the English-translated economic reports for November 2025 generated by the three methods. The following shows the economic report generated by Method A.

> Incidents and policy changes around the world are having a significant impact on the economy. A robbery in the Philippines, the suspension of SNAP benefits in the United States, and China's request to refrain from overseas travel have negatively affected consumption and the tourism industry. In addition, the shutdown of U.S. government agencies and information leaks caused by cyberattacks have created uncertainty in economic activity. Meanwhile, amid increasing investment in AI and renewable energy, there are also efforts aimed at promoting economic growth.

Although this report is factually accurate as supported by the input news articles, it may not be worth including "a robbery in the Philippines" in the report, since there was only one such news article in the input. Also, it is unclear whether "the suspension of SNAP benefits" has a significant impact on the Japanese economy.

Next, Method B identified $k'(= 3)$ topics as instructed and generated the following topic labels (in italics), summaries, and impact scores in parentheses, followed by the overall report. Note that, to save space, the topic summaries show only the snippets.

1) *U.S. Government Shutdown*: As the shutdown of U.S. government agencies continues, concerns are rising over flight reductions and disruptions to administrative functions. . . . (−4.84)
2) *China's Pressure on Japan*: The Chinese government's call for its citizens to refrain from traveling to Japan has dealt a significant blow to Japan's tourism industry and related sectors. . . . (−4.65)
3) *Suspension of U.S. Food Assistance*: The U.S. government has indicated its intention to suspend full benefit payments under the Supplemental Nutrition Assistance Program

(SNAP), raising concerns about reduced consumption among low-income households. ... (-1.73)

Amid growing global economic uncertainty, the shutdown of U.S. government agencies and mounting pressure from China on Japan are exerting significant effects. In particular, the suspension of U.S. food assistance and China's request to refrain from travel to Japan have dampened consumption and heightened uncertainty about the economic outlook. In response, countries are exploring economic countermeasures while seeking to strengthen stable supply systems and growth strategies.

Similar to Method A, the output was factually accurate, but it was not evident whether "the suspension of U.S. food assistance" has a substantial impact on the Japanese economy. Consequently, the evaluator judged that the topic was not sufficiently important to be included in the k' major topics.

Lastly, the output of Method C is shown below in the same manner.

1) *International Trade and Economic Policy*: ... International trade frictions, such as U.S. tariff hikes and China's suspension of imports, may have negative effects on the Japanese economy. ... (0.476)
2) *Revitalization of Tourism and Transportation*: ... the redevelopment of tourist destinations, the introduction of new technologies, and the hosting of events are having positive effects on regional economies. ... (0.435)
3) *Cultural and Economic Trends*: ... Cultural events and the activities of well-known individuals generally have positive effects, contributing to the revitalization of regional economies and the tourism industry. ... (0.427)

Developments in international trade, tourism, and culture are exerting multifaceted effects on the Japanese economy. In international trade, concerns persist over U.S.-China trade tensions, while Japanese companies continue to expand overseas. The revitalization of tourism and transportation is stimulating regional economies, and cultural events are supporting the tourism industry. These elements interact with one another, collectively producing complex impacts on Japan's overall economic conditions.

While the report is generally adequate as an economic report, it was not apparent what "Cultural events and the activities of well-known individuals" in the third topic referred to within the source news articles, which resulted in lower scores on the *Economic Significance* and *Factual Accuracy* metrics.

D. Results

Table I shows the results of the manual evaluation of the k' selected topics for Methods B and C, where statistically significant differences at $p < 0.05$ are marked with an asterisk (*) based on Wilcoxon signed-rank tests. Note that we used the Wilcoxon test instead of a t-test because the evaluation scores were measured on an ordinal Likert scale and the normality assumption required for the t-test cannot be guaranteed due to the small sample size (i.e., 12 months).

TABLE I. MANUAL EVALUATION OF THE k' SELECTED TOPICS.

Month	Significance		Distinctiveness		Score Validity	
	B	C	B	C	B	C
1	3	3	5	2	5	5
2	2	3	5	5	3	5
3	3	5	5	5	4	5
4	2	4	5	2	5	5
5	3	3	5	5	4	5
6	4	4	5	5	4	4
7	4	3	5	5	3	3
8	5	5	5	4	5	5
9	5	5	5	5	3	5
10	5	4	5	4	5	5
11	4	4	5	4	5	5
12	5	5	5	5	4	5
Avg.	3.75	4.00	5.00*	4.25	4.17	4.75*

Both Methods B and C performed not particularly well in *Economic Significance*, indicating that some economically significant events were not included. For instance, Sanae Takaichi was designated as the Prime Minister of Japan in October 2025 and more than 10% of the input news articles that month contained her name. Due to her economic and foreign policies, which differ from the previous prime minister's, the event should have had a significant economic impact, but it was not selected as a major topic by either method. More work is needed to identify economically significant topics. As for *Distinctiveness*, Method B worked better than Method C, even though the latter employed a separate topic model for topic identification. To improve Method C, it would be beneficial to incorporate a redundancy measure when choosing k' topics. On the other hand, Method C was rated higher in *Impact Score Validity* than Method B, indicating the strength of the topic-sentiment aggregation.

Then, Table II shows the results of the manual evaluation of the generated reports. All methods showed high *Economic Adequacy*, demonstrating the benefit of an advanced large-scale LLM in logical inference. As for *Factual Accuracy*, while the differences were not statistically significant, Method C showed lower scores on average than the other two. For instance, in the January economic report by Method C, it was stated that "China's market growth was a positive factor". However, the actual input articles contained "deflation concerns", "weak domestic demand", and "economic slowdown" in the Chinese economy, leading the evaluator to judge the statement as factually inaccurate. The particular statement was based on a news article, "China's BYD reports a 41% increase in new vehicle sales to 4.27 million units in 2024, maintaining strong growth." Controlling the influence of each news article on the final report remains an issue to be addressed in future work.

Next, Table III compares Methods B and C using the *Diversity* measure defined in Eq. (1). Consistent with the *Distinctiveness* measure in Table II, Method B performed better than Method C, partly validating the quality of the manual evaluation. In fact, *Distinctiveness* and *Diversity* scores of Method C show a positive correlation ($R = 0.59$). Note that

TABLE II. MANUAL EVALUATION OF THE GENERATED ECONOMIC REPORTS.

Month	Factual Accuracy			Economic Adequacy		
	A	B	C	A	B	C
1	5	5	2	5	5	4
2	4	4	5	4	5	5
3	5	5	5	5	5	5
4	5	5	5	5	5	5
5	5	5	3	5	5	5
6	5	5	3	5	5	5
7	3	3	3	5	5	5
8	5	5	5	5	5	5
9	5	5	5	5	5	5
10	5	5	5	5	5	5
11	5	5	3	5	5	5
12	5	5	5	5	5	5
Avg.	4.75	4.75	4.08	4.92	5.00	4.92

TABLE III. DIVERSITY OF THE k' SELECTED TOPICS.

Month	B	C
1	0.609	0.490
2	0.537	0.614
3	0.758	0.757
4	0.781	0.356
5	0.773	0.567
6	0.704	0.564
7	0.643	0.559
8	0.669	0.460
9	0.803	0.595
10	0.676	0.349
11	0.545	0.462
12	0.579	0.447
Avg.	0.673	0.518

the correlation coefficient for Method B cannot be calculated, as it has straight fives for *Distinctiveness*.

Table IV compares the API cost for generating the monthly economic reports across the three methods. Method C uses far fewer input tokens (about 33% of Methods A and B) because it first performs topic analysis locally and sends only the news headlines for the k' major topics to the LLM. As a result, the total cost of Method C is approximately 39% of the others on average. The API cost is likely negligible in practice, because only one report is generated per month and the resulting report can be stored and provided to all users of the system. However, we plan to extend the functionality to generate reports tailored to a particular event or topic of interest to individual users [30]. For instance, some users may be interested in the influence of international affairs or U.S. tariff policies on the economy. In such cases, the system would select only news articles relevant to the topic and generate a more focused summary on the fly. The total cost would then scale with the number of users and their specific needs, making cost differences more important.

Finally, Table V shows the processing time for each method, where the processing time for topic analysis and the API call are separately reported for Method C. On average, Method A was the fastest due to its simplicity, whereas Method C took about 7.5 times longer than Method A and 4.5 times longer than Method B. To be used as the on-the-fly report generation illustrated above, the processing time needs to be shortened, particularly for Method C.

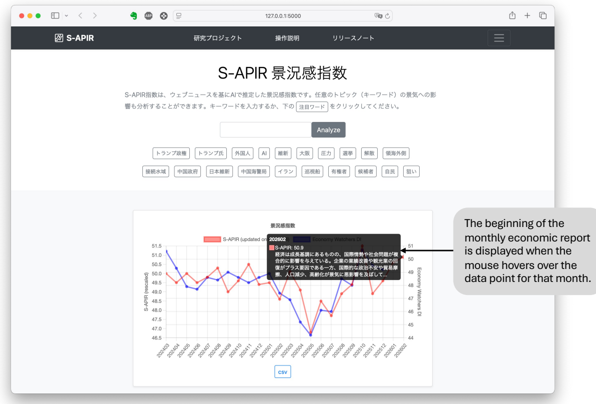

Figure 6. A screenshot of our BSI nowcasting system with a snippet of a monthly economic report.

In summary, all methods generated largely accurate and adequate reports, although Method C scored somewhat lower than the other two in *Factual Accuracy*, with no statistically significant difference. With respect to the identified major topics, Method B identified more distinctive topics than Method C, whereas the economic impact scores were found to be more valid for Method C than for Method B. In terms of API cost, Method C has an advantage over Method B; however, it requires approximately 4.5 times as much processing time. Considering these factors, a hybrid approach that combines Methods B and C may provide the best overall balance. Specifically, we first identify k distinctive topics using an LLM on a smaller subset of news articles, thereby reducing cost and processing time. We then categorize the full set of news articles locally under the identified topics. In this way, topic-level sentiment scores can also be aggregated locally to yield more appropriate estimates of economic impact.

V. IMPLEMENTATION

We designed and implemented the functionality to provide the generated report and the details of the constituent topics for Method B or C in our BSI nowcasting system. Figure 6 shows the chart of BSI values and the beginning of a monthly report for February 2026.

The chart incorporates an on-click event handler, and if the user wishes to look deeper into the month, they click the respective part of the chart and a new Bootstrap modal display appears on top of the page as illustrated in Figure 7. The modal display provides the full text of the generated economic report for the clicked month, the economic impact scores of the major topics as a bar chart, and the labels and summaries of the topics.

VI. CONCLUSIONS AND FUTURE WORK

This paper investigated methods for generating concise yet informative economic reports to enhance the interpretability of a news-based BSI. We proposed and compared the following three methods:

TABLE IV. COMPARISON OF API COST.

Month	# of input tokens			# of output tokens			Cost (US$)		
	A	B	C	A	B	C	A	B	C
1	39056	39485	13078	134	559	695	0.099	0.104	0.040
2	39210	39639	13248	148	573	695	0.100	0.105	0.040
3	39362	39791	13310	175	527	721	0.100	0.105	0.040
4	39343	39772	13277	162	593	767	0.100	0.105	0.041
5	39323	39752	13524	166	556	728	0.100	0.105	0.041
6	39881	40310	13446	171	574	763	0.101	0.107	0.041
7	39590	40019	13089	162	555	741	0.101	0.106	0.040
8	39784	40213	13260	133	618	738	0.101	0.107	0.041
9	39633	40062	13350	149	541	737	0.101	0.106	0.041
10	39828	40257	13226	155	567	695	0.101	0.106	0.040
11	39522	39951	13305	151	552	675	0.100	0.105	0.040
12	39340	39769	13435	151	522	716	0.100	0.105	0.041
Avg.	39489	39918	13296	155	561	723	0.100	0.105	0.040

TABLE V. COMPARISON OF PROCESSING TIME.

Month	A	B	C		
			Topic	API	Sum
1	5.9	8.3	22.7	25.3	48.0
2	5.5	8.9	20.9	22.1	43.0
3	5.0	8.0	20.4	30.0	50.5
4	6.7	11.0	20.9	28.8	49.7
5	12.6	8.2	19.4	23.8	43.2
6	5.5	8.7	19.5	26.7	46.2
7	4.6	19.7	19.3	26.4	45.7
8	5.7	10.8	20.7	21.3	42.0
9	4.6	10.0	19.9	25.8	45.6
10	6.2	9.4	20.3	21.0	41.3
11	4.8	7.6	19.7	18.9	38.7
12	4.4	7.7	21.1	20.7	41.9
Avg.	6.0	9.8	20.4	24.2	44.6

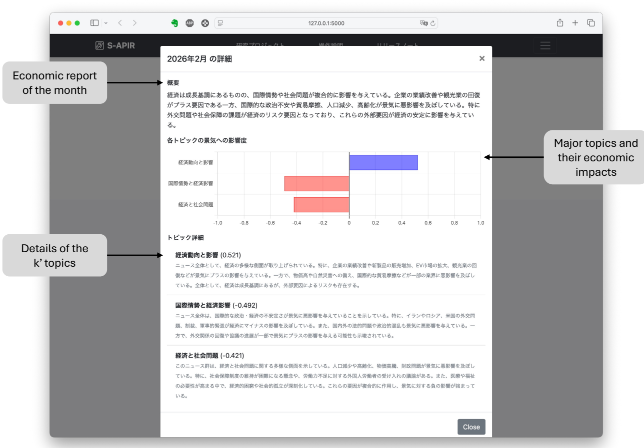

Figure 7. A screenshot of the modal display with a detailed monthly report.

- Method A uses an LLM to summarize input news headlines with high economic impact scores without considering topics.
- Method B also uses an LLM, but with a more detailed prompt to (i) categorize the input into topics and (ii) select important topics to generate an economic report.
- Method C leverages a neural topic model to identify representative, economically influential topics, which are then provided to an LLM (along with supplementary information) to generate an economic report.

We conducted empirical experiments using one year of online news data and examined both the strengths and limitations of the proposed methods. Despite incorporating a separate topic model, Method C was less effective than Method B in identifying distinctive topics. This limitation could potentially be addressed by explicitly enforcing diversity during topic selection. On the other hand, the economic impact scores of the identified topics were found to be more appropriate with Method C than with Method B. Regarding the content of the generated economic reports, the outputs were generally acceptable across all methods, although Method C tended to receive lower scores in factual accuracy. In terms of computational cost, the API cost of Method C was approximately 40% of that of the other methods; however, due to the additional topic-modeling stage, it required about 4.5 times as much processing time as Method B. Combining the strengths of Methods B and C may improve topic identification and the quality of the resulting economic reports. Additionally, we implemented functionality in our BSI nowcasting system to display the generated reports and their constituent topics, which may facilitate the collection of user feedback.

A limitation of this study is that the manual evaluation was conducted by a single expert, which may introduce subjective bias into the assessment. Additionally, the scope of the input data is constrained. Due to the excessive length of full-text news articles, our model relies exclusively on news headlines rather than the full content of the news articles. Future work may include integrating the strengths of Methods B and C, employing more efficient long-context processing techniques to incorporate full texts, and involving multiple evaluators to enhance the robustness and reliability of the evaluation results.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI #22K12280 and MEXT Japan.

APPENDIX A
INSTRUCTION TO EVALUATOR

A. Purpose of the Evaluation

The purpose of this evaluation is to assess the quality of economic reports automatically generated from one month of news articles. The evaluation should be based not on subjective preference, but on the validity of the generated content in light of the provided news articles.

B. Materials to Be Evaluated

Evaluate the reports generated by Methods A, B, and C. These reports are stored in separate files. The news articles used as input for report generation (1,000 articles per month) are also provided.

C. Evaluation Procedure

1) Review the major events of the month using the following sources:
   - Jiji Press: https://www.jijiphoto.jp/resource/html/ext/news/year/2025/index.html
   - Monthly Economic Report by the Cabinet Office: https://www5.cao.go.jp/keizai3/getsurei/getsurei-index.html
2) For each metric described next, assign a score on a five-point scale and record the results in a provided sheet. If a score of 1 or 2 is assigned, briefly state the reason.

D. Evaluation of the Three Selected Topics

This item applies only to Methods B and C. Judge based on the labels, importance scores, and summaries of Topics 1–3. The monthly economic report itself should not be evaluated here.

- Economic Significance
  Evaluation Criteria:
  - Whether major topics that significantly affected the economy during the month are included
  - Whether important topics are omitted
  - Whether minor issues are selected as major topics
  - Refer to the input articles if needed

  Scoring:
  - 5: All three topics are economically important
  - 4: All three topics are economically relevant
  - 3: Difficult to judge
  - 2: Some topics are not economically relevant
  - 1: None of the topics are economically relevant
- Distinctiveness
  Evaluation Criteria:
  - Whether the three topics substantially overlap in content (economic relevance is not required)
  - It is not necessary to check the input articles

  Scoring:
  - 5: All three topics are clearly distinct
  - 4: Two topics are somewhat similar but distinguishable
  - 3: Difficult to judge
  - 2: Two topics are nearly identical
  - 1: All three topics are nearly identical
- Impact Score Validity
  Evaluation Criteria:
  - Whether the relative magnitudes of the three economic impact scores (the numerical values following the topic labels) are reasonable
  - No comparison with other months or methods is required; judge only within the same month
  - It is not necessary to check the input articles; judge based on the topic descriptions

  Scoring:
  - 5: Relative magnitudes are appropriate
  - 4: Generally appropriate
  - 3: Difficult to judge
  - 2: Relative magnitudes appear questionable
  - 1: Relative magnitudes are clearly inappropriate

E. Evaluation of the Generated Economic Report

For this item, evaluate only the “economic report for the month” for each method and do *not* evaluate the labels or summaries of Topics 1–3 here.

- Factual Accuracy
  Evaluation Criteria:
  - Whether information not contained in the news articles has been added
  - Whether numerical values or events are incorrectly described
  - Whether any misleading expressions are present
  - Refer to the input articles if needed

  Scoring:
  - 5: No factual errors are identified
  - 4: Only minor issues are present
  - 3: Some parts are difficult to judge
  - 2: One clear factual error is present
  - 1: Multiple factual errors are present
- Economic Adequacy
  Evaluation Criteria:
  - Whether the interpretation of economic causal relationships is reasonable
  - Whether there are logical gaps
  - Whether there are unclear or unintelligible interpretations
  - Whether the report is natural in an economic context
  - Do not refer to the input articles; evaluate whether the logic of the report itself is coherent. Even if factual errors are present, focus solely on the internal economic reasoning.

  Scoring:
  - 5: Highly plausible
  - 4: Generally plausible
  - 3: Difficult to judge

– 2: Contains implausible interpretations
– 1: Not plausible at all

F. Notes for Evaluators

- Do not evaluate based on stylistic preference.
- Evaluate the content rather than writing quality.
- If uncertain, select 3.
- If assigning a score of 1 or 2, or if any notable issue is identified, record it briefly.

References

[1] K. Seki, "How to interpret an economic index? Generating reports with topic sentiment analysis," in *Proceedings of the Thirteenth International Conference on Building and Exploring Web Based Environments (WEB 2025)*, 2025, pp. 9–10.

[2] Cabinet Office, Government of Japan, *Economy watchers survey*, https://www5.cao.go.jp/keizai3/watcher-e/index-e.html, Accessed: Feb. 7, 2026.

[3] Bank of Japan, *Tankan*, https://www.boj.or.jp/en/statistics/tk/index.htm, Accessed: Feb. 7, 2026.

[4] K. Seki, Y. Ikuta, and Y. Matsubayashi, "News-based business sentiment and its properties as an economic index," *Information Processing & Management*, vol. 59, no. 2, 102795, 2022.

[5] A. H. Shapiro, M. Sudhof, and D. J. Wilson, "Measuring news sentiment," *Journal of Econometrics*, vol. 228, no. 2, pp. 221–243, 2022. DOI: https://doi.org/10.1016/j.jeconom.2020.07.053.

[6] Y. Yamamoto and Y. Matsuo, "Sentiment summarization of financial reports by LSTM RNN model with the Japan Economic Watcher Survey Data," in *Proceedings of the 30th JSAI*, In Japanese, 2016.

[7] K. Seki, "Refining sentiment predictions: Obtaining an unbiased business sentiment index from Japanese newspapers," *International Journal of Asian Language Processing*, vol. 33, no. 2, 2350015, 2023.

[8] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, Second. Oxford University Press, 2012.

[9] M.-C. Bieri, "Assessing economic sentiment with newspaper text indices: Evidence from Switzerland," *Swiss Journal of Economics and Statistics*, vol. 161, 10, 2025. DOI: 10.1186/s41937-025-00139-4.

[10] E. Kalamara, A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia, "Making text count: Economic forecasting using newspaper text," *Journal of Applied Econometrics*, vol. 37, no. 5, pp. 896–919, 2022. DOI: 10.1002/jae.2907.

[11] J. Ashwin, E. Kalamara, and L. Saiz, "Nowcasting Euro area GDP with news sentiment: A tale of two crises," *Journal of Applied Econometrics*, vol. 39, no. 5, pp. 887–905, 2024. DOI: 10.1002/jae.3057.

[12] L. Barbaglia, S. Consoli, and S. Manzan, "Forecasting with economic news," *Journal of Business & Economic Statistics*, vol. 41, no. 3, pp. 708–719, 2023. DOI: 10.1080/07350015.2022.2060988.

[13] W. Hu, X. Zhang, and Y. Ren, "Generating financial reports from macro news via multiple edits neural networks," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2020, pp. 667–682. DOI: 10.1007/978-3-030-67664-3_40.

[14] S. Yan, "Disentangled variational topic inference for topic-accurate financial report generation," in *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, 2022, pp. 18–24. DOI: 10.18653/v1/2022.finnlp-1.3.

[15] Z. Zhang *et al.*, "Enhancing multi-document summarization with cross-document graph-based information extraction," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023, pp. 1696–1707. DOI: 10.18653/v1/2023.eacl-main.124.

[16] N. K. Shukla *et al.*, "Generative AI approach to distributed summarization of financial narratives," in *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 2872–2876.

[17] X. Yang, S. Zang, Y. Ren, D. Peng, and Z. Wen, *Evaluating large language models on financial report summarization: An empirical study*, 2024. arXiv: 2411.06852 [cs.CL].

[18] G. Assis *et al.*, "Language models for automated market commentary from corporate disclosures," in *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*, 2025, pp. 727–735. DOI: 10.1145/3768292.3770438.

[19] P. Cui and L. Hu, "Topic-guided abstractive multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1463–1472.

[20] L. Dong, M. N. Satpute, W. Wu, and D. Du, "Two-phase multidocument summarization through content-attention-based subtopic detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1379–1392, 2021.

[21] C. Li, A. Xu, S. Joty, and G. Carenini, "Topic-guided reinforcement learning with LLMs for enhancing multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 12 395–12 412.

[22] B. Warner *et al.*, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," *arXiv preprint arXiv:2412.13663*, 2024.

[23] The Sankei Shimbun, *The Sankei Shimbun*, https://www.sankei.com, Accessed: Feb. 7, 2026.

[24] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002.

[25] X. Wu, T. T. Nguyen, D. C. Zhang, W. Y. Wang, and A. T. Luu, "FASTopic: Pretrained Transformer is a fast, adaptive, stable, and transferable topic model," in *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[26] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100–108.

[27] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[29] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022. arXiv: 2203.05794 [cs.CL].

[30] X. Zhang, Q. Wei, Q. Song, and P. Zhang, "TOMDS (topic-oriented multi-document summarization): Enabling personalized customization of multi-document summaries," *Applied Sciences*, vol. 14, no. 5, 1880, 2024.