2023 European Conference on Computing in Construction
40th International CIB W78 Conference
Heraklion, Crete, Greece
July 10-12, 2023

# IDENTIFYING THE FACTORS OF COUNTRY RISK FLUCTUATION FROM NEWS TEXT DATA USING NATURAL LANGUAGE PROCESSING

Sehwan Chung[1], Jungyeon Kim[1], Seokho Chi[1,2], and Du Yon Kim[3]
[1]Seoul National University, Seoul, Korea, Republic of (South Korea)
[2]Institute of Construction and Environmental Engineering, Seoul, Korea, Republic of (South Korea)
[3]Kyungil University, Gyeongsan, Korea, Republic of (South Korea)

## Abstract

Due to the uncontrollability of country-level risk, project participants like contractors should keep monitoring the current situations of the host country. News articles might be good sources to extract issues preceding country risk fluctuation. This study proposes a framework to recognize risk-related issues from news content. S-BERT vectorizes news texts, and DBSCAN groups them into several topics. Topics from different timespans are compared to distinguish risky issues. Preliminary results show the proposed framework can extract specific topics from news articles. It is expected that participants of international projects utilize the framework to recognize the situations of the host country.

## Introduction

In international construction projects, the changes in the business environments of a host country, known as host country risks, are critical to the successful delivery of the projects. For example, it has been shown that the host country's political, social, economic, and environmental conditions affect the cost of an international construction project (Zhu et al., 2020). Due to the uncontrollable nature of country-level risk events, it is essential for the project participants to keep monitoring current issues of the host country and to take timely actions when they identify any precursor of risk events. Consequently, participants of international construction projects, especially the general contractors, are devoted to monitoring current issues in the host country, which may develop into risk events.

Because news articles published in the host country describe diverse topics, a collection of news might be a promising source from which researchers and practitioners can extract issues related to the fluctuation of country risk. However, due to a large number of news articles published daily, it is impractical to monitor every news article and track recurring events that may develop into risks.

Therefore, there is a need to develop a methodology (1) to collect a large number of news articles in an automated way, (2) to extract important topics from the news data, and (3) to identify issues potentially related to the fluctuation of host country risk. To the authors' best knowledge, however, few existing studies have explored the potential of news articles as a source to identify the factors of country risk fluctuation.

To address this knowledge gap, this study proposes a framework to identify the issues potentially related to the fluctuation of country risk. In the proposed framework, a natural language processing (NLP) method is used to represent the news text into numeric vectors while preserving the meaning of the original news content. In addition, Density-based Spatial Clustering of Applications with Noise (DBSCAN) is used to group news into topics. News topics from different timespans are then compared to distinguish the common and distinctive issues.

## Literature Review

Many studies in the construction domain have investigated the effects of country risks on the performance of projects from various perspectives. One group of studies focused on country risks involved in a joint venture (JV) of construction firms, a common approach for a multinational firm to enter a foreign construction market (Bing et al., 1999).

Ozorhon et al. (2007) investigated the impacts of host country conditions on the performance of international JVs. They used a questionnaire survey to measure the effects of host country conditions on the performances of international JVs. In their study, host country conditions were categorized into political risks (e.g., inconsistency in policies), macroeconomic conditions (e.g., fluctuations in inflation or foreign exchange rates), the strength of the legal system, and the relations between JV and the host government. Their results showed an indirect impact of host country conditions on JV performances via the factors related to a specific project. For example, political stability and macroeconomic conditions were found to affect project-related factors such as the completeness of the contract or the project definition, which consequently affects performance. Similarly, Hwang et al. (2017) used a questionnaire survey to evaluate the criticality of risk factors for international construction JVs between Singapore and developing countries. They categorized 29 risks, derived from the literature review, into country-level, market-level, and project-level risks. Survey results indicated that political instability was the most critical risk factor, followed by project budget overrun, corruption, and uncertain market demand. With a specific focus on cultural similarities/differences, Ozorhon et al. (2008) examined the effects of the cultural gap between JV partners and between a JV partner and the host country through a questionnaire survey. While the results showed that the differences in organizational culture significantly impacted JV performances, the effect of cultural differences between the host country and a JV partner was insignificant.

Another research focus was the quantitative analysis using empirical data on international construction projects. Lee et al. (2015) analyzed the effects of host countries on

international construction projects by Korean contractors. Their results showed that the situations of host countries affected the project performances to some extent. Such effects were indirect, more significant in developing countries than developed countries, and more critical for large construction firms than small or medium-sized firms. In line with the goal of their previous study, Lee and Han (2017) analyzed the database of Korean contractors' past project performances in 32 overseas countries. They classified the countries by four factors: business environment, market opportunity, the possibility of project success, and market experience. As a result, the 32 countries were classified into three main groups: (1) countries with high market opportunity and moderate business environment, (2) countries with a highly favorable business environment but moderate market opportunity, and (3) countries with low market opportunity and hostile business environment. Last, Li et al. (2020) focused on social risks from the public of host countries involved in railway construction projects by Chinese contractors. Based on the grounded theory methodology, their study collected news with the keyword "Belt and Road Portal" and developed a theory on the public resistance to railway construction projects. Their results revealed how public resistance emerged and affected the delivery of projects. They also suggested risk management strategies for international contractors to deal with such social risks.

However, the existing studies are limited because they heavily relied on questionnaire surveys or macro-level datasets to obtain the data for analyses. In many studies, questionnaire surveys were used to extract knowledge from the domain experts. Consequently, the scope of risk identification and assessment was limited to the experience and knowledge of the experts, which is not exhaustive enough to cover the broad aspect of country risk. In the studies using macro-level databases, country-level risk factors were already identified through a literature review, and the usage of empirical data was limited to test their well-defined but fixed assumptions on the relationship between risk factors and macro-level indices. In reality, there are too many risk events to be covered by a fixed set of hypotheses.

With the advance of data analytics, it would be possible to identify the factors of country risk fluctuation from the extensive dataset of actual events, such as news text data. However, the existing studies are limited in exploring the potential of extensive news datasets as the data source for risk factor identification. To overcome the limitations, the objective of this study is to propose and validate a data-driven methodology to identify factors of country risk fluctuation, especially from news text data.

## Methodology

Figure 1 presents the overall methodology of this study, which consists of six main procedures as follows.

- Collecting open-source data
- Preprocessing data
- Embedding news text
- Segmenting timespans
- Clustering news topics from text
- Extracting issues from news topics

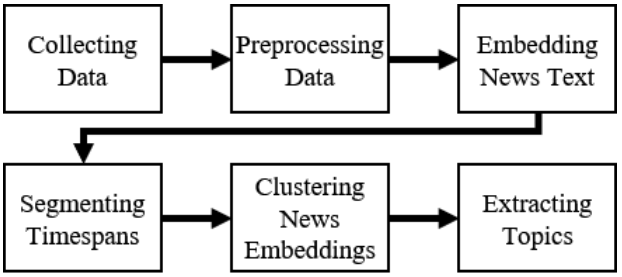


*Figure 1: Overall Research Methodology*

The outputs of the proposed framework are the issues extracted from the news articles of a country of interest (i.e., the host country). Therefore, it can address the problem stated in the previous sections. First, an extensive collection of news articles is automatically collected. Then, this framework applies (1) Sentence Bidirectional Encoder Representations from Transformers (S-BERT) and (2) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to group the contents of the news articles into several topics. Consequently, the proposed framework can provide significant issues from the extensive collection of news articles, with the minimum human effort of collecting and analyzing the news.

To the authors' best knowledge, it is the first attempt in the construction domain that S-BERT is used to extract issues from the news corpus. In addition, the preliminary study by the authors showed that news clustering based on S-BERT could provide more informative topics than the results by existing methods, such as topic modeling with Latent Dirichlet Allocation (LDA), which will be discussed later.

### Collecting Open-source Data

The proposed methodology collects two types of data: news text and risk-related indices. News texts are used as sources from which risk-related issues are extracted. Risk-related indices refer to the quantitative indicators of the country risk fluctuations.

Obtaining news data is challenging because of the tremendous amount of daily streaming news data. To tackle this issue, the proposed methodology applies web scraping for the automated collection of news text data from online news publishers. Web scraping is the method of decoding a web page written in Hypertext Markup Language (HTML) and gathering specific contents from the decoded web page (Nicolas et al., 2021).

A web scraping program only works when the address of a web page to be scraped is provided. Thus, it is essential to build a list of all the web addresses of the online news articles to be collected. A web scraping guideline for each website, called 'robots.txt,' can be used for this task. This text file specifies allowable actions (e.g., accessing a publicly open web page) and disallowable actions (e.g., ex-

ploiting the search engine of the website) in the website. In addition, a typical 'robots.txt' file provides the sitemap that contains the list of all accessible web pages of the website. A web scraping program can utilize the sitemap to gather the web addresses of all web pages (i.e., online news articles) under the selected news website. Once the web address list is completed, a simple rule-based web scraping program would work fine since all the web pages in the same news publisher usually have the same HTML structure.

Risk-related indices refer to the indicators of the fluctuations of country risk. The country risk index, such as country risk ratings by S&P, Moody's, or World Bank, can be used as the country risk indicator. However, the country risk ratings tend to be published quarterly or yearly. Thus, it would be hard to extract issues related to the fluctuation of the indicators because too many issues would happen over a relatively long period, interacting with each other in a complex way.

Instead of using such high-level indicators of country risk, utilizing specific factors of host country risk is more practical. For example, an economic/financial index such as material price, foreign exchange rate, or oil price can be used as the indicator of country risk fluctuation. These indices are created in a relatively short time, like monthly, daily, or even in real-time, and also have high volatility, making it a plausible approach to use news text to extract the issues related to the fluctuation of these indices. In practice, many economic and financial indices are available online and thus can be obtained easily.

**Preprocessing Data**

Most demand for data preprocessing emerges from the quality problems in the collected news text data. There may be several data quality problems. For instance, missing or erroneous values in news data (e.g., title, publication date, body text), different structures for the same element (e.g., publication date written in 2023-07-10 or July 10, 2023), and unnecessary text elements (e.g., advertisements) may exist. Therefore, data exploration needs to be conducted on the collected dataset to find any data quality problem in the news text data and improve the data quality.

**Embedding News Text Data**

Text embedding is the task of representing an input text into a numeric vector. The purpose of text embedding is not just to convert text data into a computer-processable format but also to conserve the meaning of the news content. In other words, text pieces (e.g., sentences) with similar meanings should be vectors with similar values.

Recently, Bidirectional Encoder Representations from Transformers (BERT) has achieved state-of-the-art performances in various NLP tasks (Devlin et al., 2019). This study uses Sentence-BERT (S-BERT), a modification of BERT, as a backbone architecture for embedding news text (Reimers and Gurevych, 2019). The core idea of S-BERT is to convert input sentences into vectors so that the output vectors represent the semantic relationship of the original sentences. If two sentences are semantically similar, the output vectors will have a high similarity measure.

Due to this characteristic, S-BERT is a suitable option for the proposed methodology. S-BERT is expected to represent the news text into vectors so that news about the same topic is located nearby in the embedding vector space. Therefore, any well-established clustering algorithm will work fine to identify the latent topics of the collected news text.

The overall architecture of S-BERT is presented in Figure 2. Given an input text, the BERT model represents each token as a vector, resulting in a sequence of vectors. A pooling layer is added to compress the results in a fixed-sized vector. S-BERT uses average pooling to maintain the most informative values from the BERT with minimum information loss.

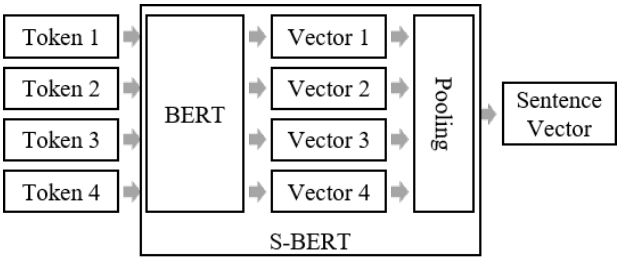


*Figure 2: Architecture of S-BERT*

Pre-training an S-BERT model is as follows. Two S-BERT models are jointly trained, as depicted in Figure 3. During the training, pairs of two input sentences are given, each sentence input into each S-BERT model, respectively. The resulting two vectors determine the semantic relationship between two input sentences. The relationship can be given as categorical values (e.g., contradiction, entailment, neutral) or numeric values (e.g., the similarity score between -1 and +1).

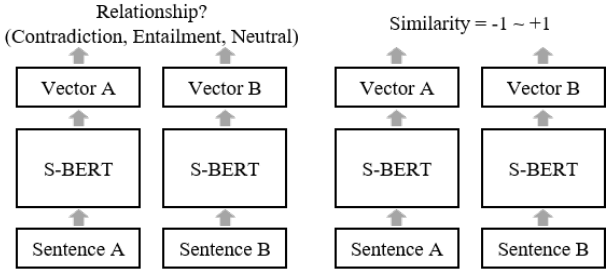


*Figure 3: Pre-training of an S-BERT Model*

**Segmenting Timespans**

The news text embeddings are segmented into several timespans by a certain period (e.g., day, week, month). Each timespan is annotated according to the direction of the risk index fluctuation. Every news within a specific timespan can be related to the changes in the risk index. In reality, of course, many topics would be irrelevant to the fluctuation of the risk index. Here, the fundamental assumption of this study is that if one topic repeatedly hap-

pens with a specific upper or lower movement of a risk index, that topic may be an issue related to such movement of the risk index.

The selection of the length of each timespan is based on the results of clustering. On the one hand, clustering all the news text without segmentation is not practical due to the massive number of news articles to be analyzed, usually exceeding hundreds of thousands. On the other hand, setting a too short period results in each timespan containing only several news articles is insufficient to draw meaningful topics.

**Clustering News Embeddings**

A clustering algorithm is applied to the news embeddings in each timespan, resulting in several clusters of news text (i.e., news topics). It should be considered that not all news articles are essential. Some news may describe daily happenings unrelated to the fluctuation of country risk. Thus, typical clustering algorithms such as k-Means or Latent Dirichlet Allocation (LDA) are not suitable for news clustering because they try to allocate every news article into at least one cluster regardless of the importance of the news.

This study uses Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to cluster the news embeddings into topic clusters (Ester et al., 1996). The core assumption of DBSCAN is that a dataset may have noisy data that should not be labeled as clusters. In addition, it also assumes that each cluster would have at least a certain number of data points. Because of these characteristics, DBSCAN is suitable for clustering news data as topics. Once an issue happens, news publishers tend to publish articles repeatedly about the issue for a certain time, which fits the DBSACN's assumption of a minimum number of data points for a cluster. A topic described only by a few news articles might not be an essential issue.

The procedures of DBSCAN are described in Figure 4. Two key hyperparameters of DBSCAN are epsilon (*eps*) and the minimum number of samples (*min_samples*). If the distance between two data points is smaller than *eps*, the two data points are directly connected. If two data points, even though they are not directly connected, are connected to another data point simultaneously, the two data points are also connected. If the number of connected data points is equal to or greater than *min_samples*, the connected data points are considered one cluster. All the other data points not belonging to a cluster are considered noises.

It should be noted that the number of clusters is not a hyperparameter in the DBSCAN algorithm. Once *eps* and *min_samples* are set, the algorithm automatically determines what points should be grouped into one cluster or classified as noises without uncertainty. Therefore, the number of clusters of DBSCAN is determined by the settings of *eps* and *min_samples*.

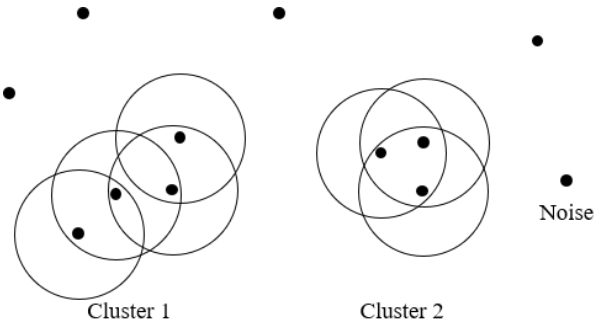


*Figure 4: Mechanism of DBSCAN*

**Extracting Issues from News Topics**

The result of the DBSCAN algorithm is the set of clusters labeled with arbitrary numbers. Therefore, it is required to extract information from each cluster that can describe the topic of the news cluster. Keywords of a cluster, such as the most frequent words in the text, can help interpret news clusters into understandable topics.

However, preliminary studies by the authors showed it was still hard to extract issues from the clustering results because too many topics exist across multiple timespans. By the nature of the DBSCAN algorithm, it is not guaranteed that cluster 1 in the first timespan is the same as cluster 1 in the second or third timespan, because the cluster label is given arbitrarily.

Therefore, this study repeatedly applies a clustering algorithm to the clustering results from the previous step to overcome this issue. Specifically, non-noise data points are selected from the clustering results, which are considered news describing important topics. Then, the DBSCAN clustering algorithm is applied to the news to figure out topics happening over multiple timespans. Combined with the fluctuation of the risk index, each cluster (i.e., news topic) is labeled as either *increase*, *decrease*, or *stable*, depending on the portion of increase/decrease news data points. For example, if many news in one cluster are from the timespans when the risk index has increased, the cluster is labeled as *increase*. It eventually represents that the news in this cluster is potentially related to the increase in the risk index.

## Results

As an example case, this study applied the proposed methodology to extract issues related to the exchange rate of United States Dollars and the Philippines Peso (USD-PHP). The foreign exchange rate, a typical risk factor of international construction projects, was chosen as the risk index to be analyzed. It significantly affects profitability and is known to be influenced by many global and local issues. The Philippines was selected as a country with relatively high country risk compared to other countries where English is an official language (e.g., the United Kingdom).

### Data Collection

News data were collected from the official website of Manila Bulletin, one of the national-wide news publishers in the Philippines. A Python-based web scraping program was developed and collected 310,875 news articles.

### Data Exploration and Preprocessing

Figure 5 shows the distribution of news published by year. News posted in 1970 must be erroneous and were removed during the data preprocessing. In addition, the bar chart shows that only 92 articles were published before 2018. After a manual investigation of the news before 2018, the authors also removed the news from the analysis dataset.

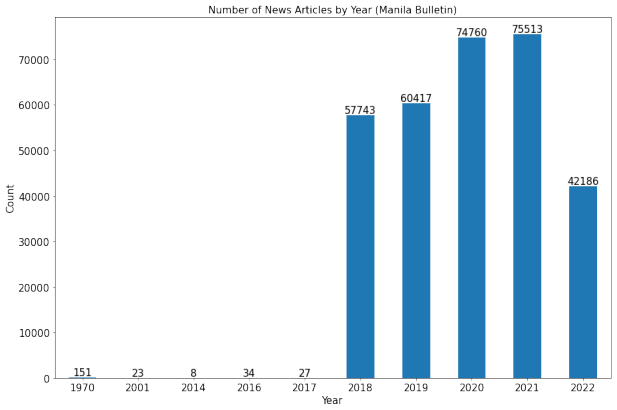


*Figure 5: Number of News Articles by Year*

### News Text Embedding

This study used the 'all-mpnet-base-v2' model for news embedding, one of the pre-trained S-BERT models (Song et al., 2020). It achieved the best average performance in sentence embeddings and semantic search tasks. Despite its high computational cost, the selected model was chosen because the authors implemented the news embedding on NVidia RTX A6000; thus, the computational cost was not a critical issue.

Since there was no quantitative or objective metric to measure the performance of news embeddings, the authors qualitatively investigated the results of news embeddings. Figure 6 shows the selected S-BERT model's visualization of all news embeddings. For example, the red boxed area in Figure 6 contains news articles about vaccines and COVID-19, which can be interpreted from the keywords of the news text. The most frequent words included "vaccine(s)," "COVID-19," "vaccination," and others. It means that news articles about the same topic (i.e., COVID-19 and vaccines) were embedded in a nearby area in the embedding space.

Further investigations on different areas (i.e., green and blue) qualitatively verified that news text was embedded as intended. For instance, news vectors within the green box were about natural disasters, with frequent words of PAGASA (Philippine Atmospheric, Geophysical and Astronomical Services Administration), "cloudy," "weather," and "thunderstorms." The blue box represents the topic of politics and diplomacy (keywords: "North Korea," "Kim," "Korean," "Trump").

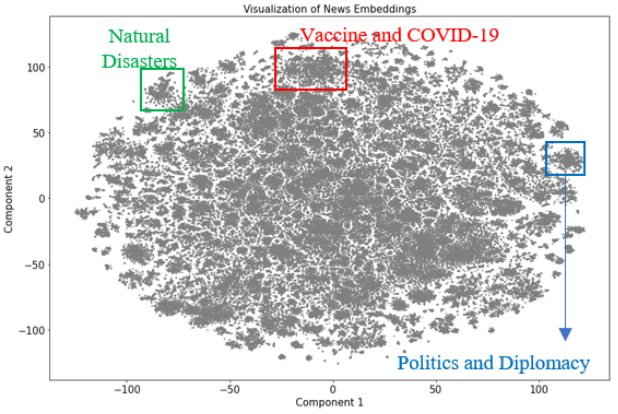


*Figure 6: Visualization of All News Embeddings*

### Timespan Segmentation

This study segmented the dataset into months, resulting in 56 timespans (i.e., 56 months) from January 2018 to August 2022. Each month was labeled according to the movement of the USD-PHP exchange rate. The months with exchange rate fluctuation exceeding a certain threshold were annotated as *increase*. Similarly, the month with exchange rate fluctuation below a certain threshold was classified as *decrease*, while the remainings were classified as *stable*. The cutoff thresholds were manually determined, as +0.5 for the *increase* cutoff and -0.5 for the *decrease* cutoff, based on the distribution of monthly differences in the USD-PHP rate (Figure 7).

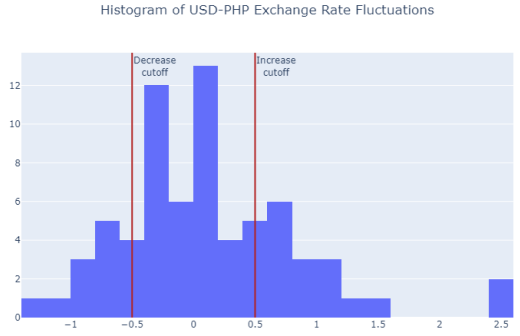


*Figure 7: Distribution of USD_PHP Rate Fluctuations*

### News Clustering

The news embeddings were clustered for each time segment by the DBSCAN algorithm. When clustering text data of each month, the hyperparameters of DBSCAN were fixed as *eps* to 0.6 and *min_samples* to 10. These parameters were empirically chosen on a trial-and-error basis, and they resulted in a balance between the number of clusters and the number of data points in each cluster.

Table 1 summarizes the clustering results. For each month, 77.56% of news data were labeled as noise by DBSCAN

on average, and there were 28 topic clusters on average. At least 13 clusters were identified by DBSCAN for every month, meaning the hyperparameter setting was appropriate for extracting meaningful topics from news.

*Table 1: Summary of Clustering Results*

| Number of Clusters | | | The Ratio of Noises(%) | | |
|---|---|---|---|---|---|
| Min. | Max. | Avg. | Min. | Max. | Avg. |
| 13 | 74 | 28.0 | 50.9 | 87.5 | 77.6 |

### Interpretation and Extraction of Issues

After identifying representative news topics of each month from the previous step, the DBSCAN algorithm was again applied to the non-noise data points to identify issues from the representative topics. In this step, hyperparameters were set to *eps*=0.6 and *min_samples*=100.

Figure 8 visualizes the results of issues extracted from important news text. A total of 22 clusters were identified, and 53.55% of the data were labeled as noises. In each cluster, every news is either from the month when the exchange rate increased or decreased. Each cluster was annotated as *increase*, *decrease*, or *stable* according to the portion of increase/decrease in the exchange rate. Within a topic, if the majority of news is linked to the increase, it is classified as *increase*. If the portion between increase and decrease was balanced, the topic was classified as neutral, representing that this topic is not related to the fluctuation of the exchange rate.

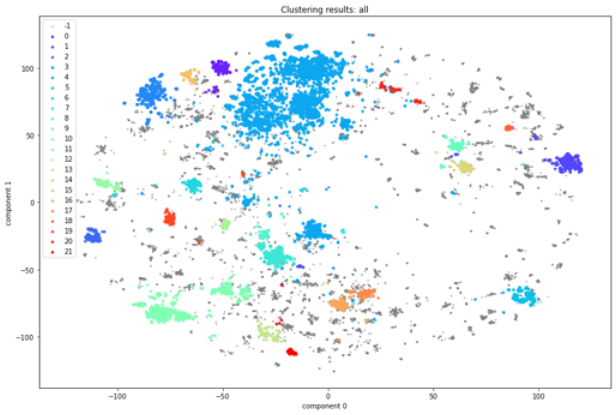


*Figure 8: Visualization of Extracted Topics*

Table2 summarizes the results of the issue extraction. For each cluster, the five most frequent keywords were identified. Based on the set of keywords, the authors manually named each topic by its primary content. For example, the first cluster (Cluster 0) includes the keywords of "Phivolcs (Philippine Institute of Volcanology and Seismology)," "earthquake," "city," "intensity," and "km." Thus it is reasonable to name this topic as 'natural disaster and earthquake.' In addition, the balance of news article numbers in the period of increase/decrease shows that the issues are not significantly related to the exchange rate fluctuations.

Eight topics were identified as potentially related to the increase of the USD-PHP exchange rate, and five were identified as potentially rate-decreasing issues. Even though it does not guarantee that the extracted issues affected the exchange rate fluctuation, it can be inferred that such issues are related to the exchange rate fluctuation. For example, an issue in the Philippine stock market might have affected the exchange rate increase (Cluster 2), or the diplomacy-related issue (Cluster 7) might be related to the exchange rate decrease.

It should be noted that it does not guarantee that this issue 'caused' the exchange rate fluctuation. There might be some unrevealed relationship between the issue and the fluctuation. Despite the limitation, this preliminary result can provide an overview of the issues happening in the country of interest and guides the construction practitioners to the starting point of country risk identification and monitoring.

## Discussion

First, the results demonstrated that the proposed framework could provide enough information to extract significant issues from news articles in a specific country. The framework successfully extracted local issues like specific disasters (e.g., earthquakes, volcano eruption) and local political issues (e.g., election, PhilHealth), besides global issues such as COVID-19 or international political issues, Methodologically, S-BERT showed its capability of vectorizing the contents of news articles. Partial investigations demonstrated that news vectors in a specific location share the same topic. Such ability may come from the power of an S-BERT model pre-trained on a vast corpus. For example, the MPNet model was trained on large-scale text corpora over 160 GB with 32 high-performance NVIDIA Tesla V100 GPUs. The rapid advancement of pre-trained language models in the NLP field would increase the performance of news text embedding, enhancing the capability of the proposed framework.

Compared to the existing topic modeling with the typical LDA algorithm, this framework provided much more informative and intuitive keywords for each topic. The authors' unpublished preliminary study revealed that the results of LDA were hard to interpret; the keywords often contained stopwords (e.g., "news," "Mr," or "said"), making it hard to guess the underlying issue of the topic extracted by LDA. In contrast, the clustering results based on the S-BERT were straightforward. With minimum knowledge of the issues in the case study country (i.e., the Philippines), the authors could extract the underlying issues of each topic by only examining the frequent keywords and the titles of several news articles in each cluster.

One advantage of the proposed framework is that it consists of several modules: a data collecting and preprocessing module, a text embedding module, a clustering module, and an interpretation module. Therefore, researchers or practitioners can change each module with an arbitrary method they want to apply. For example, the DBSCAN clustering algorithm can be replaced with other clustering

*Table 2: Clustering Results with Monthly Fluctuations of USD-PHP Foreign Exchange (ForEx)*

| Cluster | No. of News | Keywords (five most frequent words) | Topic Name | No. of Increase | No. of Decrease | ForEx Movement |
|---|---|---|---|---|---|---|
| 0 | 286 | Phivolcs, earthquake, city, intensity, km | Disaster (earthquake) | 136 | 150 | Stable |
| 1 | 686 | North, Korea, Kim, Trump, Korean | Diplomacy | 495 | 191 | Increase |
| 2 | 296 | market, percent, billion, local, investors | Financial (stock) | 206 | 90 | Increase |
| 3 | 903 | PAGASA, may, weather, thunderstorms, country | Weather, Disaster | 560 | 340 | Stable |
| 4 | 6,324 | COVID-19, cases, government, vaccine, City | COVID-19 | 3,337 | 2,987 | Stable |
| 5 | 352 | points, PBA, Ginebra, Magnolia, TNT | Sports (basketball) | 199 | 153 | Stable |
| 6 | 199 | rice, NFA, farmers, price, country | Food industry | 89 | 110 | Stable |
| 7 | 697 | China, President, Chinese, Duterte, Sea | Diplomacy | 245 | 452 | Decrease |
| 8 | 154 | Kuwait, workers, OFWs, Filipino, President | Social | 106 | 48 | Increase |
| 9 | 218 | China, trade, US, Trump, Chinese | Diplomacy | 108 | 110 | Stable |
| 10 | 1,192 | Drug, police, Police, shabu, City | Social | 656 | 536 | Stable |
| 11 | 177 | per, oil, prices, price, liter | Economy (oil price) | 96 | 81 | Stable |
| 12 | 102 | budget, House, Senate, President, 2019 | Politics (budget) | 67 | 35 | Increase |
| 13 | 118 | NPA, Army, Infantry, troops, government | Politics | 82 | 36 | Increase |
| 14 | 224 | Hong, Kong, China, police, protesters | Politics | 104 | 120 | Stable |
| 15 | 151 | Taal, Volcano, Phivolcs, volcanic, Alert | Disaster (Volcano) | 127 | 24 | Increase |
| 16 | 220 | Duterte, Mayor, President, Sara, Davao | Politics (election) | 170 | 50 | Increase |
| 17 | 109 | Robredo, President, Leni, Vice, presidential | Politics (election) | 78 | 31 | Increase |
| 18 | 104 | Saudi, Khashoggi, Arabia, consulate, Trump | General issue | 0 | 104 | Decrease |
| 19 | 251 | PhilHealth, President, corruption, Morales, Senate | Politics (PhilHealth) | 2 | 249 | Decrease |
| 20 | 72 | cases, deaths, country, virus, coronavirus | COVID-19 | 0 | 72 | Decrease |
| 21 | 149 | law, bill, terrorism, President, Anti-Terrorism | Politics | 18 | 131 | Decrease |

algorithms, such as k-means or other state-of-the-art methods. The MPNet model was used in our study, but any pre-trained language model can be used for news text embedding. For those who already secured the news text data to be analyzed, the data collection process can be omitted. In setting the hyperparameters of the DBSCAN, the authors applied a trial-and-error basis. For example, setting too small *eps* resulted in all data points being classified as noises, while setting too large *eps* grouped all data points into one cluster. Without a quantitative metric of the clustering performance, the authors had to determine the balance between the number of clusters and the ratio of noises heuristically. Therefore, it is not guaranteed that the settings of hyperparameters in this study are optimal. It leads to the need to build a dataset of news articles labeled by human annotators as a basis for quantitative validation of the clustering results.

## Conclusions

This study proposed an NLP framework to extract issues related to the host country risk from news text by applying web scraping, text embedding based on Sentence-BERT, and the clustering algorithm of DBSCAN. Preliminary results showed that the proposed framework could effectively represent news articles as vectors while preserving their contents with the S-BERT method. In addition, the DBSCAN method can extract distinctive topics from news articles while filtering out routine news data. Any participant in international construction projects could utilize the proposed framework to recognize the important issues that happened or are happening in the host country. Furthermore, past and current issues are expected to be used for informed decision-making on managing host country risk. The limitation of this study includes that the interpretation of the clustering results still relies on manual investigation. Even though the proposed methodology suggested how to automate data collection procedures, news text embedding, and clustering of the embeddings, users still need to investigate each cluster and figure out the main topic of the cluster. It limits the applicability of the proposed methodology because the manual interpretation of news clusters requires prior knowledge of the user on the issues in the host country. Another limitation is the lack of quantitative validation of the results, as discussed in the previous section.

Future studies should be thus conducted on automating the interpretation of the resulting news clusters and providing processed information that a user with limited prior knowledge can understand. For example, automated and unsupervised topic labeling can be one option that aims to label each topic, making it easier to grasp the meaning of each news topic. In addition, ground truth (i.e., news articles labeled by human annotators) should be established to base the thorough and quantitative validation of the proposed framework.

## Acknowledgments

This research was supported by the National Research Foundation of Korea, grant number [NRF-2021R1F1A1062701]. This research was supported by

Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0008475, Development Program for Smart Digital Engineering Specialist).

## References

Bing, L., Tiong, R. L.-K., Fan, W. W., and Chew, D. A.-S. (1999). Risk management in international construction joint ventures. Journal of Construction Engineering and Management, 125:277–284.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Association for Computational Linguistics.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.

Hwang, B.-G., Zhao, X., and Chin, E. W. Y. (2017). International construction joint ventures between singapore and developing countries risk assessment and allocation preferences. Engineering, Construction and Architectural Management, 24:209–228.

Lee, K.-W. and Han, S. H. (2017). Quantitative analysis for country classification in the construction industry. Journal of Management in Engineering, 33:04017014.

Lee, K.-W., Han, S. H., Park, H., and Jeong, H. D. (2015). Empirical analysis of host-country effects in the international construction market: An industry-level approach. Journal of Construction Engineering and Management, 142:04015092.

Li, H., Zhong, Y., and Fan, C. (2020). Reducing the social risks of transnational railway construction: a discussion on the formation mechanism of host country people's coping behaviors. Engineering, Construction and Architectural Management, 28:1657–1682.

Nicolas, C., Kim, J., and Chi, S. (2021). Natural language processing-based characterization of top-down communication in smart cities for enhancing citizen alignment. Sustainable Cities and Society, 66:102674.

Ozorhon, B., Arditi, D., Dikmen, I., and Birgonul, M. T. (2007). Effect of host country and project conditions in international construction joint ventures. International Journal of Project Management, 25:799–806.

Ozorhon, B., Arditi, D., Dikmen, I., and Birgonul, M. T. (2008). Implications of culture in the performance of international construction joint ventures. Journal of Construction Engineering and Management, 134:361–370.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. pages 3982–3992. Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding.

Zhu, F., Hu, H., Xu, F., and Tang, N. (2020). Predicting the impact of country-related risks on cost overrun for overseas infrastructure projects. Journal of Construction Engineering and Management, 147:04020166.