

Highlights

Deep-learning Approach for Economic Index Construction in ICT industry: A Case Study of Korea

Dong Eun Min,Nathaniel Kang,Yonghun Cho,Joonyeon Choi,Jongho Im

- Research highlights item 1
- Research highlights item 2
- Research highlights item 3

Deep-learning Approach for Economic Index Construction in ICT industry: A Case Study of Korea^{*,**}

Dong Eun Min¹, Nathaniel Kang¹, Yonghun Cho², Joonyeon Choi³ and Jongho Im^{*,1}

Department of Statistics and Data Science, Yonsei University, Seoul, 03726, South Korea

Department of Software, Sejong University, Seoul, 02748, South Korea

ARTICLE INFO

Keywords:
Data fusion
Deep learning
Economic index
forecasting
KoBERT

ABSTRACT

The ICT industry refers to the industry of “information and communications technology”. In South Korea, the ICT industry plays an important role in the economy, possessing 12.8% of the GDP of South Korea. Due to recent unforeseen incidents, it has become a challenge to accurately predict the trend of the ICT industry. This study aims to generate an economic index through a deep learning approach that integrates Natural Language Processing (NLP) models and image clustering techniques. We introduce a 4-step protocol to generate an economic index to evaluate economic performance. First, systematic sampling was applied to produce a balanced sample of ICT industry news for each period. Second, feature engineering techniques developed by KoBERT were applied to generate two types of scores - relevance scores and sentiment scores. Third, textual data were transformed into joint plot images for visualization and then grouped into different clusters based on news categories. In the end, Multi-criteria Decision Analysis (MCDA) is applied to generate the final economic index.

1. Introduction

The ICT industry refers to the industry of “information and communications technology,” which includes manufacturing and services industries whose products are related to information processing and communication by electronic means, such as semiconductors, laptops, and mobile phones. In South Korea, the ICT industry plays an important role in the economy, possessing 12.8% of the GDP of South Korea. As the demand for computer, communication, and consumer electronic products (3C) significantly increased along with the COVID-19 pandemic in 2020, fluctuations in ICT exports occurred. According to the ICT Industry Trend Report by the Export-Import Bank of Korea, exports of semiconductors in the first quarter of 2021 increased by 13.2% compared to the same period of the previous year, due to a strong increase in demand for mobile devices [11]. Also, for displays, exports expanded by 18.8% due to the increased demand for OLED panels. In particular, OLED exports increased by 32% compared to the same period last year [10].

Due to recent unforeseen incidents, it has become a challenge to accurately predict the trend of the ICT industry. Since most economic data are released with a lag and are subsequently revised, both forecasting and assessing current-quarter conditions are important tasks for central banks [8, 12]. For example, structured macroeconomic data such as monthly employment rates released by both government agencies usually have a significant lag of 1 month [3].

Due to the publication delay, it poses a challenge for econometrics and government policymakers to measure real-time economic performance.


In order to effectively monitor the real-time macroeconomic conditions, economists at central banks shifted to using unstructured data such as textual news to distill relevant information [3, 4]. Hence, unstructured data can be applied to address the nowcasting issue. Since unstructured data such as text data from social media are generated every day, textual data can help reflect and monitor real-time events. Nevertheless, handling unstructured data poses another great challenge due to its complexity. This paper aims to propose a novel framework to predict an economic index in the ICT industry by integrating sampling techniques, Korean Bidirectional Encoder Representations from Transformers (KoBERT), and Multiple-criteria Decision Analysis (MCDA).

This study introduces a 4-step framework to generate an economic index. First, systematic sampling was applied to produce a balanced sample of ICT industry news for each period. This process ensures that the training data consists of the same amount of news articles for each time period and media. Second, feature engineering techniques developed by KoBERT were applied to generate two types of scores - relevance scores and sentiment scores. Third, news articles were transformed into images for visualization and then grouped into different clusters. In the end, Multi-criteria Decision Analysis (MCDA) model is applied to generate the economic index.

The main contributions of this paper are summarized as follows:

- Understanding unstructured data such as textual news data can be served as a valuable source to assess the economic outlook and capture the economic activities, such as production, consumption, investment, etc. By

*Corresponding author

 demin@yonsei.ac.kr (D.E. Min); natekang@yonsei.ac.kr (N. Kang); yhcho8587@yonsei.ac.kr (Y. Cho); zoon@sejong.ac.kr (J. Choi); ijh38@yonsei.ac.kr (J. Im)

ORCID(s): 0000-0002-1773-0419 (D.E. Min); 0000-0002-6305-7341 (N. Kang); 0000-0003-1635-2345 (Y. Cho); 0000-0002-6604-5944 (J. Choi); 0000-0001-8362-4756 (J. Im)

exploiting information from unstructured data, we can create useful insights.

- Providing both the reasoning behind and the purposes of adopting text-based data as opposed to traditional economic indicators. Also, this paper analyzes the advantages and limitations of using soft indicators and hard indicators respectively.
- Proposing a four-step framework to generate an economic index by utilizing textual news data. The four-step framework includes systematic sampling, Natural Language Processing (NLP), image clustering, and obtaining an estimated economic index through Multi-criteria Decision Analysis (MCDA).
- Implementing the proposed framework in the ICT industry in South Korea to validate the performance. Based on the experiment results, the economic index estimated by our four-step framework outperforms other current economic indices and it is able to measure ICT-related economic activities by successfully capturing the fluctuation of the economic activities.

2. Related Work

Standardized hard indicators (e.g. unemployment rate, exports volume) and soft indicators (e.g. business surveys, consumer spending level) have been widely used to estimate and predict key macroeconomic outcomes [14, 22, 25]. However, hard indicators have posed several structural problems [14, 25]. One of the problems is that hard indicators are calculated periodically (either quarterly or annually) at a certain time, so it is difficult to reflect the real-time fluctuations of the economy. If there exists a major event that affects the overall national economy, such as the financial crisis of 2008 or COVID-19, hard indicators are not able to accurately reflect these economic movements [28].

The above-stated problem can be overcome by soft indicators. However, soft indicators also have some limitations. Since data for soft indicators are collected by a traditional survey method, it is possible that sampling errors may occur depending on the response rate. Furthermore, constructing the indicator itself may become difficult if the response rate is extremely low. For example, the Economic Sentiment Indicator (ESI) survey has become more difficult to be conducted in the United Kingdom and Italy during the nationwide COVID-19 lockdown [1]. Furthermore, it is not easy to find a survey pool designed for highly-specialized industries such as the ICT industry given the time and cost issues.

In order to address the problem, the use of text-based unstructured information can be considered as an alternative to surveys, and text-based data has been proven to have strong predictive power in several studies [7]. Unstructured data has several advantages. First, not only is it possible to understand the general public's perception of the industry through social media, but it is also possible to collect data

from various industry reports and industry news articles. Accordingly, the use of unstructured data is receiving attention in the macroeconomic area and financial industry [5].

Among various types of text-based information, this study selects online newspaper articles as the only source to predict the economic performance in the ICT industry for the following reasons. First, newspaper articles are published every day, so textual news is able to provide real-time economic information. Second, newspaper articles are carefully written by professional reporters or editors, so the quality is more stable than regular social media posts. Furthermore, most textual news belong to its corresponding categories such as political, entertainment, and economic news, allowing us to select the proper dataset in the experiment.

3. Proposed Framework

Our proposed framework includes three major components: KoBERT, Image Clustering, and MCDA. The architecture of the proposed method is illustrated in Fig. 1.

3.1. KoBERT

In order to determine if a news article belongs to ICT-industry news and measure its sentimental level, deep learning approaches such as neural networks can be utilized. There are several innovative NLP models available such as Doc2Vec [13], GloVe [21], and BERT [6]. However, these models are not able to handle the complexity of the Korean-language context. Since the Korean language differs greatly in structure from English or other languages, English-language-based NLP models might lead to poor performance. To address this language issue, this study selects KoBERT¹ as our core NLP technique to perform sentimental analysis by computing relevance score and sentiment score.

KoBERT is a model designed by T-Brain which aimed to improve the performance of BERT (Bidirectional Encoder Representations for Transformers) for Korean textual data. Even though BERT already includes a multilingual model which can be used immediately for multiple languages, KoBERT is particularly designed for implementation to the Korean language.

KoBERT consists of two structural features: the data-based tokenization technique, and ring-reduce based distributed learning method. By tokenization, the model can reflect the characteristics of Korean, which has irregular changes compared to languages such as English or French. Also, ring-reduce based distributed learning allows the model to learn large amounts of data in a short period of time. Specifically, KoBERT can quickly learn more than 1 billion sentences on multiple machines through distributed learning technology.

In this study, we construct a classification model with sentiment features and combine it with ensemble learning. The customized KoBERT model can be divided into three stages, including input, process, and output. In the input

¹SK Telecom, <https://github.com/SKTBrian/KoBERT>

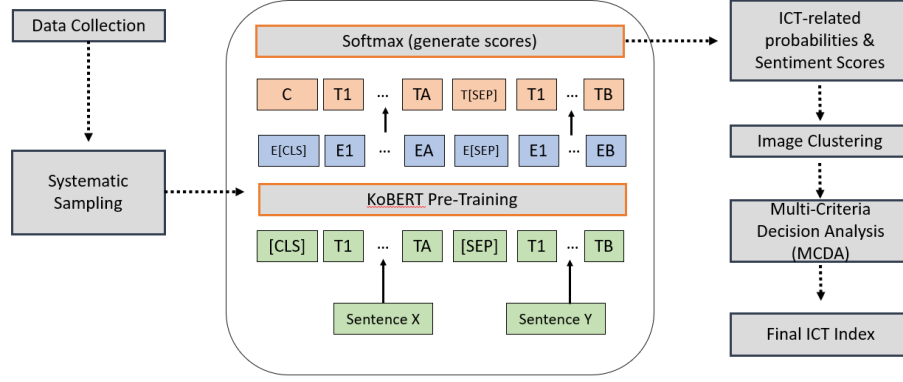


Fig. 1: Our proposed framework consists of 3-major components: KoBERT, Image Clustering, and MCDA

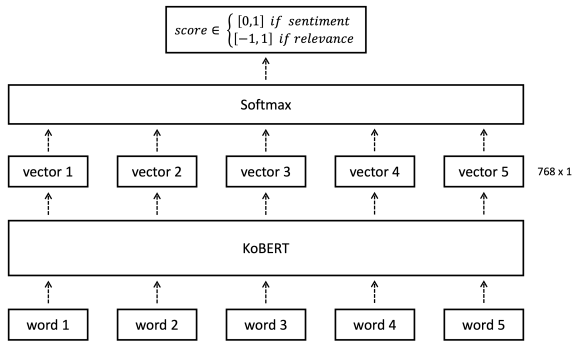


Fig. 2: Structure of the KoBERT model

stage, the model performs pre-classification to identify the news into 3 categories: ICT-related, neutral, and Non-ICT-related. In the process stage, input features are fed to the pre-trained KoBERT model to tune the KoBERT model. After tuning the KoBERT model for the classification task, the features of text sentiments are integrated into the model through an ensemble learning method to boost the model performance [15]. In the output stage, the probability of ICT-related news is calculated through the model and then the sigmoid activation function is applied to obtain relevance scores and sentiment scores. Fig. 2 demonstrates the KoBERT architecture.

3.2. Image Clustering

After KoBERT generates two scores - relevance scores and sentiment scores, textual news articles will be transformed into 2-dimensional images. The image is a 2-dimensional heatmap where the x-axis refers to the ICT relevance scores, and the y-axis refers to sentiment scores. An example plot is illustrated in Fig. 3.

The main purpose of transforming news articles into images is for data visualization. Visualizing the images allows us to access the performance of the clustering algorithm and inspect the clustering results. Next, we perform feature extraction with VGG16 [23], a type of convolutional neural networks (CNN) to extract image features from each

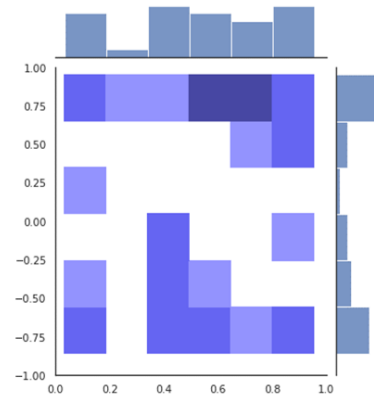


Fig. 3: Example of a joint plot

image. VGG16 is a CNN consisting of 16 layers with 13 convolutional layers, 5 pooling layers, and 3 fully connected layers with 1 softmax function. The VGG16 architecture is illustrated in Fig. 4.

This study adopts a customized VGG16 model with the last two layers and the final output is a 4096-dimensional vector. Since every image is the form of a high-dimensional vector, Principal Component Analysis (PCA) is applied for dimension reduction by projecting high-dimensional data into a low dimension. Through PCA, the variability of the data can be expressed with a relatively small number of variables.

After extracting important features for every news article, K-means clustering algorithm [16] is applied to cluster the images into several sub-groups. K-means clustering algorithm searches for the optimal clusters by repeatedly assigning each observed value to a cluster and computing new cluster centroids by distance, with a given number of clusters. Lastly, the elbow method [27] is used to determine the optimal number of clusters.

3.3. Multi-criteria Decision Analysis (MCDA)

Multi-Criteria Decision Analysis (MCDA) is a widely-used decision-making tool that can be applied to many complex decisions in several fields. Belton and Stewart [2]

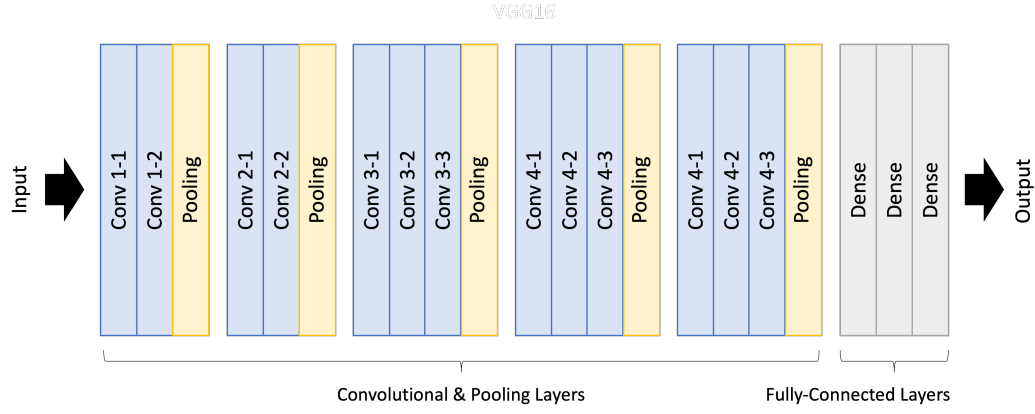


Fig. 4: Model Architecture of VGG16

define MCDA as “an umbrella term to describe a collection of formal approaches, which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter.” In this study, six MCDA methods were applied to determine the cluster ranking: TOPSIS [9, 30], VIKOR [31], ARAS [32], EDAS [18], MAIRCA [20], and MARCOS [24]. The main goal of MCDA is to make the decision-making process easier and the results more consistent by explicitly considering multiple criteria and structuring complex decision problems [19]. Inspired by Thokala et al. [26], we developed the following key steps to rank the clusters: defining the problem, selecting and developing criteria, weighting criteria, and scoring alternatives.

In selecting and developing criteria, our decision candidates are the clusters previously generated. The decision criterion is the category of each article. Based on these candidates and criteria, we proposed three components - decision matrix, weight matrix, and criterion positivity. First, we construct a decision matrix based on the news category ratio of each cluster.

Second, the weight matrix was obtained by calculating the entropy of each category and ranking them. The entropy formula is expressed as in Eq. 1. After we calculate the entropy of each category, we rank the categories in descending order and use these ranks to compute the final weight matrix. The weights of each candidate are normalized so that the sum equals to 1, as in Eq. 2.

$$H(x) = - \sum_{k=1}^K P(x_k) \log P(x_k) \quad (1)$$

$$w_i = r_i / \sum_{k=1}^K r_k, \quad (2)$$

where r_i = rank of i th category.

Lastly, we determined the positivity and negativity of each category with the ratio of news articles related to the ICT industry. If the ratio of articles relevant to the ICT

industry exceeds 0.5, the category is considered a positive category and vice versa.

After ranking all candidates by iterating every MCDA algorithm, final cluster weights $c_i^{(t)}$ are calculated in the reverse order of the cluster rankings. For example, a cluster with the highest weight will be ranked first. That is, the final cluster weight $c_i^{(t)}$ can be expressed as $c_i^{(t)} = 7 - k_i$, where $k_i = 1, 2, \dots, j$ and j indicates the number of the cluster. In the end, the MCDA method selection is based on two criteria - correlation between the estimated index and actual ICT-GDP, and the volatility of the estimated index.

3.4. Final Index Prediction

Computing the final index requires 4 major steps. First, we re-scale the relevance score $v_{ij}^{(t)}$ and sentiment score $s_{ij}^{(t)}$ for each sentence $j (j = 1, \dots, m_i)$, where t indicates the time period and i indicates the i 'th article which consists of m_i sentences. Re-scaled relevance score $v_{ij}^{*(t)}$ is calculated by squaring the existing relevance score $v_{ij}^{(t)}$ and then normalizing it so that the sum of the ICT relevance scores equals to 1 for each article. The re-scaled sentiment score $s_{ij}^{*(t)}$ is calculated such that it follows a truncated log-normal distribution. Re-scaled relevance scores and sentiment scores are expressed in Eq. 3 and Eq. 4 respectively.

$$v_{ij}^{*(t)} = \frac{v_{ij}^{2(t)}}{\sum_{j=1}^{m_i} v_{ij}^{2(t)}}, \quad (3)$$

$$s_{ij}^{*(t)} = F^{-1}\left(\frac{s_{ij}^{(t)} + 1}{2}\right), \quad (4)$$

where F is a CDF of truncated log-normal distribution with $\text{meanlog} = 100$, $\text{sdlog} = 100$, $\text{min} = 0$, and $\text{max} = 200$.

Second, we aggregate the re-scaled relevance scores and sentiment scores for each article. This article score is calculated by multiplying the previously re-scaled relevance scores by the sentiment scores for each sentence and then summing them up for each article.

$$S_i^{(t)} = \sum_{j=1}^{m_i} v_{ij}^{*(t)} s_{ij}^{*(t)} \quad (5)$$

Third, we rank each cluster based on its relevance to ICT industry by applying every Multi-Criteria Decision Analysis (MCDA) method. In the experiment, we selected VIKOR as the final MCDA method. After ranking, each cluster is given a cluster weight of $c_i^{(t)} = 7 - k$, where k is the rank of the cluster. Then, we obtain the estimated ICT index $I_t^{(b)}$ by multiplying the article score $S_i^{(t)}$ by its cluster weight and calculating the average.

$$I_t^{(b)} = \sum_{i=1}^{n_t} c_i^{*(t)} S_i^{(t)}; c_i^{*(t)} = \frac{c_i^{(t)}}{\sum_{i=1}^{n_t} c_i^{(t)}} \quad (6)$$

Lastly, we repeat the three steps done above $B = 30$ times and get the final ICT index I by calculating the median of the 30 indices obtained.

$$I = \text{median}\{I_t^{(1)}, I_t^{(2)}, \dots, I_t^{(B)}\} \quad (7)$$

4. Experiment

To evaluate the performance of our proposed 4-step framework, we performed the experiment on real-world news data from Naver News², a news service of the Korean internet giant Naver Corporation. Naver News is one of the most comprehensive news sources in South Korea, providing access to global and domestic content.

4.1. Data Processing

This study aims to generate an economic index to evaluate the performance of the Korean ICT industry. Therefore, a rigorous sampling process must be conducted in order to ensure appropriate data is pulled from the database. After the target population was determined, there were about 49 million news articles collected from 2010 to 2021. Next, the raw data underwent three major sampling tests such as topic test, category test, and media test. These three tests serve as a sampling process to eliminate the data which is not related to the ICT industry. For example, articles in categories such as "weather" or "fashion" were eliminated from the category test.

After filtering the raw data through these tests, about 6 million articles remained. Then, we performed systematic sampling to the remaining articles. Systematic sampling is a probability sampling method that selects members from the population at a regular interval. The approach was proposed by Yates in 1948 [29]. Unlike other simple random sampling techniques, systematic sampling eliminates the probability of cluster selection and lowers the chance of obtaining contaminated data. Therefore, by using the systematic sampling

²Naver News, <https://news.naver.com/>

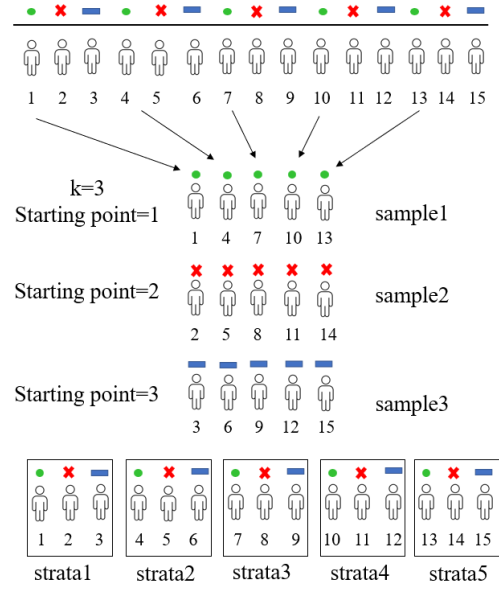


Fig. 5: Implicit stratification in systematic sampling method

method, we can obtain samples that reflect the characteristics of the population. To perform systematic sampling, a target population has to be pre-determined before selecting the corresponding participants. Theoretically, systematic sampling determines the sampling interval k which is obtained by dividing the population size by the desired sample size. After determining k , the systematic sampling method selects the first member from 1 to k randomly and then selects every k -th member of the target population. Fig.5 demonstrates how systematic sampling performs at $k = 3$.

Through the process above, we obtain 3 different datasets. First, 4,000 articles with 10 to 60 sentences were selected as training data for KoBERT. Second, 14,400 articles were used as prediction data for constructing monthly and quarter indices. Lastly, 31,350 articles were used as prediction data for weekly indices. The dataset consists of the news ID, publication date, title, and sentences of the article. In particular, news ID has a unique value as a combination of year and month, and the sentence in a news article is used as an analysis target. For the details of the training dataset and variable description, refer to appendix A1 and A2.

4.2. Relevance & Sentiment Score Prediction

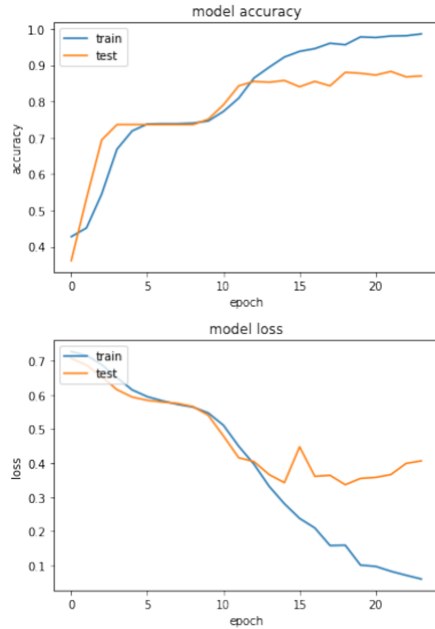
Using the data constructed above, we trained two KoBERT models: The relevance prediction model and the sentiment prediction model. In the training process, the data was divided into a training set (80%) and testing set (20%). Since the data for KoBERT consisted of 4,000 articles, the training set consisted of 3,200 articles and the testing set consisted of 800 articles. Fig. 6 shows the model loss and accuracy for each epoch. It is clear that the training accuracy of the model increases as the training procedure continues, but the test accuracy stopped improving after the 13th epoch.

After training the two models, each sentence in a news article was embedded into a 2-dimensional vector consisting

Table 1

KoBERT Experiment Example: relevance score and sentiment score for each article

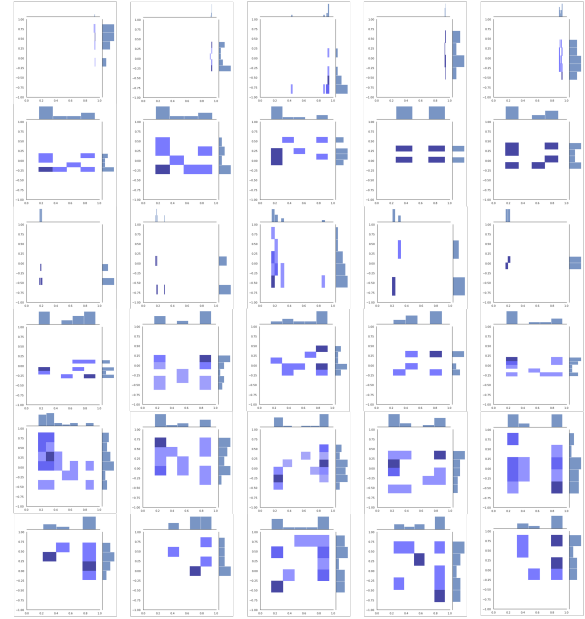
Article ID	Date	Sentence	Relevance score	Sentiment score
44516805	2010-01-03	Woori Bank increased stable investment products, while decreasing products focusing on bonds.	0.007	-0.326
44517505	2010-01-04	Although the export volume of the first step was only \$10,000, the company has gained confidence and sales have increased by 10%.	0.343	0.941
44519905	2021-11-19	While producing prices have risen throughout the year, consumer prices are also expected to rise.	0.125	-0.024

**Fig. 6:** KoBERT results

of a relevance score and a sentiment score. The relevance score identifies how the news article is closely related to ICT industry, and the sentiment score assesses the tone of the sentence on a spectrum of positive to negative. The relevance score was calculated in the range of $[0, 1]$, whereas the sentiment score was calculated in the range of $[-1, 1]$. If the sentence is predicted to be positive, its score is closer to 1, while for the opposite, its score is closer to -1. Tab 1 shows an excerpt from the KoBERT training results.

4.3. Plot Image Clustering

Here, each news article in the data was transformed into a plot with a 4096-dimensional vector. Next, Principal Component Analysis (PCA) was applied to reduce the dimensionality from 4096 to 14 dimensions. According to our observation, 14 principal components are able to explain 95% of the overall variance. Lastly, we used k-means clustering to group every article which was represented as a heatmap to different clusters and applied the elbow method to determine the optimal number of clusters. From the experimental result, the optimal value of k is 6.

**Fig. 7:** Image clustering results. Each row represents an example of plots that belong to a distinct cluster.

The image clustering results are summarized in Fig. 7. Images listed on each row represent its corresponding cluster. For instance, the images on the first row belong to the first cluster, those on the second row belong to the second cluster, and so on. Based on our visualized representation of the k-means clustering results, images with similar features were grouped in the same cluster. This implies that we successfully group the news articles into several clusters based on their distinct features.

4.4. Multi-criteria Decision Analysis (MCDA)

After image clustering, we ranked each cluster by its relevance to the ICT industry. The rank and weight of each cluster were determined by Multi-Criteria Decision Analysis (MCDA). In the experiment, each news article belongs to one of the following 15 categories: IT, finance, general economy, computer, industry, stock market, global economy, telecommunications, internet, science, games, mobile, local economy, security & hacking, and small and medium-sized enterprises (SME). The news category can be considered a critical metric to evaluate ICT industry relevance. For example, news from IT, computer, and telecommunication

Table 2

Decision Matrix for MCDA

Cluster	IT	Games ...	Computer	Telecommunication
1	0.207	0.085 ...	0.027	0.041
2	0.031	0.002 ...	0.002	0.002
3	0.333	0.027 ...	0.043	0.098
4	0.120	0.039 ...	0.009	0.021
5	0.231	0.089 ...	0.030	0.047
6	0.087	0.009 ...	0.009	0.012

categories are highly related to the ICT industry. On the other hand, news from the finance category are less likely to have a strong correlation with the ICT industry. Hence, the news category was used as the decision criterion and then we constructed a decision matrix with the size of 6×15 , which is the number of clusters and the number of news categories respectively. The elements of the matrix are defined as the category ratio of each cluster. Tab. 2 shows an example of the decision matrix.

Next, the weight matrix was obtained by calculating the entropy of each category. Here, the entropy is defined with the ratio of how much each category is related to the ICT industry. Since there are two scenarios - being relevant to the ICT industry and being irrelevant to the ICT industry, we define K as 2 and $P(x_1)$ is defined as the proportion of articles that are related to the ICT industry, while $P(x_2) = 1 - P(x_1)$. For example, for the "IT" category, $P(x_1)$ is 0.574 and $P(x_2)$ is 0.426. Then we can apply the values obtained previously to Eq. 1 to calculate the entropy of each news category. Since the entropy becomes larger when the difference of $P(x_1)$ and $P(x_2)$ gets small, we rank the categories in descending order and use these ranks as weights after normalizing them so that the sum equals to 1. Tab. A4 demonstrates an example of the weight matrix.

Finally, we determined the positivity and negativity of each category with the ratio of news articles related to the ICT industry from table A3. If the ratio of articles relevant to the ICT industry exceeds 0.5, the category is considered a positive category, and otherwise, it is considered a negative category.

Through generating constructed decision matrix, weight matrix, and positivity for each category, we ranked all clusters using various MCDA methods such as TOPSIS, VIKOR, ARAS, EDAS, MAIRCA, and MARCOS. Then, the final cluster weights were calculated in the reverse order of the cluster rankings.

After obtaining the cluster weights, the final index was calculated by re-scaling the relevance scores and sentiment scores of each sentence and aggregating these scores for each article. Refer to Tab. 3 which provides an example of the obtained article scores.

By multiplying the article scores with the cluster weights obtained previously, the index for one iteration is computed. After repeating this process 30 times, the median of the 30 indices generated is selected as the final index. With the aim of selecting the best MCDA method, we compared each

Table 3

Example of Article Scores

Article ID	Date	Category	Cluster	Score
13	2017-01-01	Finance	1	100.576
14	2010-06-01	Finance	1	21.871
19	2015-12-01	IT	2	170.216
21	2018-04-01	IT	3	90.694
...
142022	2016-01-31	General Economy	1	136.207

Table 4

Volatility of Each MCDA Method

MCDA	Volatility (Quarterly)	Volatility (Monthly)
TOPSIS	114.754	259.309
VIKOR	63.821	187.319
ARAS	109.987	262.089
EDAS	102.597	257.367
MAIRCA	72.059	191.546
MARCOS	118.352	266.600

MCDA method based on the volatility of the indices and further compared the estimated index with actual GDP of the ICT industry to inspect if they follow a similar trend. In the end, we concluded that VIKOR outperformed other MCDCA methods as VIKOR achieved the lowest index volatility and follows the trend of actual ICT-GDP. Therefore, we chose VIKOR as the optimal MCDA method, and then we further compared the result from VIKOR with actual ICT-GDP and other economic indices used in South Korea to reflect the economic status of the ICT industry.

4.5. Experimental Results

In this section, we validate whether our proposed index is able to reflect the economic fluctuations of the ICT industry in South Korea. Here, we compared our index (ICT-NSI) with actual ICT-GDP values and the ICT Business Survey Index (ICT-BSI), a survey-based indicator compiled by the Bank of Korea. The ICT-BSI is constructed by conducting surveys about the business conditions of the current month and outlooks of the following month on business executives of the industry. To compare the forecasting power of our index and ICT-BSI, we first compare the quarterly ICT-GDP growth rate with the quarter-on-quarter change of our proposed index and ICT-BSI respectively. Then, we compare the forecasting results of ICT-GDP growth rates using our index and ICT-BSI.

Fig. 8 shows the comparison between the growth rate of actual GDP with our proposed index and ICT-BSI respectively. From this figure, we can see that while ICT-BSI is not able to reflect the fluctuations of ICT-GDP, our proposed index is able to capture the fluctuations so that the proposed index follows a similar trend as that of ICT-GDP.

Fig. 9 shows the forecasting results of ICT-GDP growth rates. Here, we compare the quarterly forecasting results using ICT-GDP and ICT-NSI with the results using ICT-GDP and ICT-BSI. The data from 2010 to 2019 are used

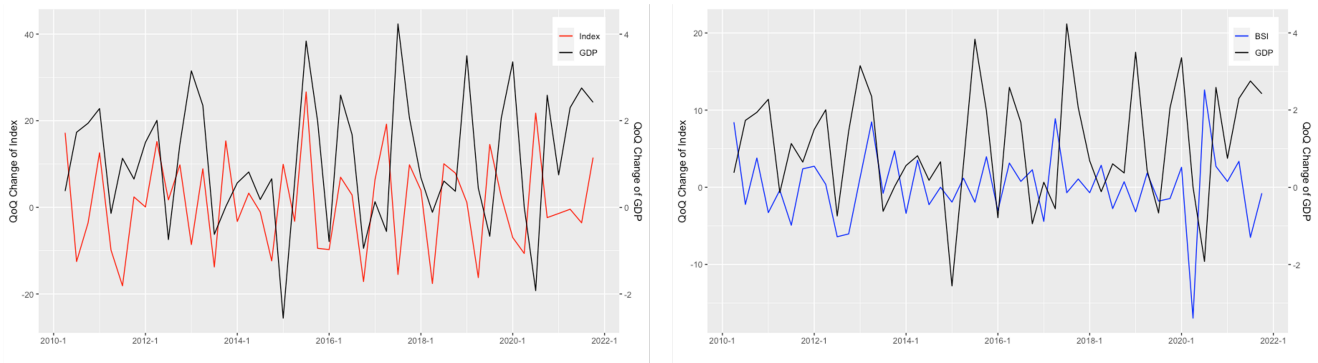


Fig. 8: Comparison of ICT-GDP Growth Rate (Quarterly) with ICT-NSI & ICT-BSI

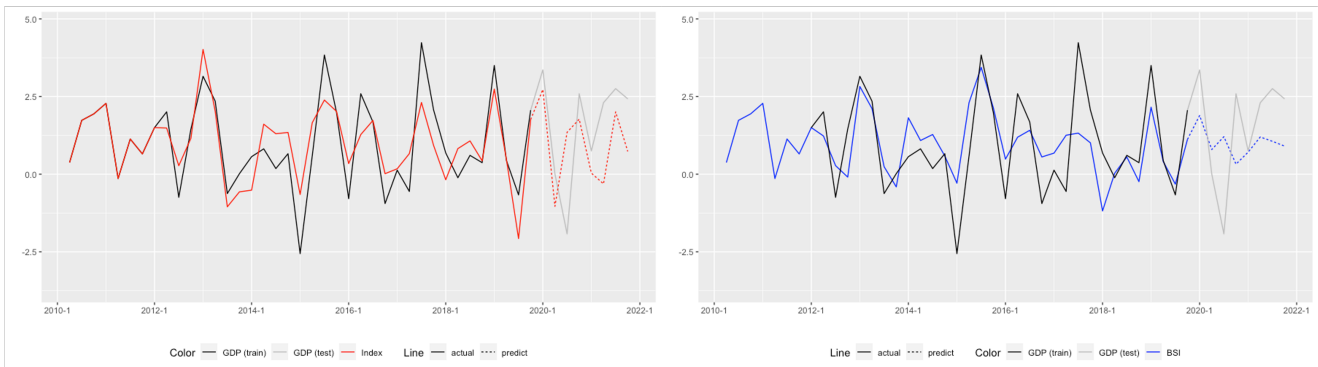


Fig. 9: Forecasting Result of ICT-GDP Growth Rate (Quarterly) using ICT-NSI or ICT-BSI. The red solid (blue solid) line represents the in-sample predictions obtained from the model that includes our index (BSI) and ICT-GDP. The dotted lines represent the out-sample predictions of each model. The black solid (black dotted) line shows the in-sample (out-sample) target variable, ICT-GDP.

Model	$RMSE_{in}$	MAE_{in}	$RMSE_{out}$	MAE_{out}
ICT-NSI	0.952	0.799	1.718	1.441
ICT-BSI	1.164	0.920	1.736	1.501

Table 5

Forecasting accuracy of ICT-GDP growth rate using ICT-NSI & ICT-BSI

for in-sample forecasting and those from 2020 to 2021 are used for out-sample forecasting. In reference to [1], we used a bi-variate mixed frequency vector autoregressive model (MF-biVAR) for forecasting. Since ICT-BSI is produced as a monthly index, while our index is provided in both monthly and quarterly terms, we aggregate the monthly ICT-BSI into quarterly values following the aggregation relationship of [17] when using it as an explanatory variable.

Tab. 5 shows the in-sample and out-sample forecasting accuracy of the two cases. Root mean squared error (RMSE) and mean absolute error (MAE) are used as accuracy metrics. For both in-sample forecasting and out-sample forecasting, we can see that ICT-NSI reflects the trend of ICT-GDP better than ICT-BSI, and therefore obtains lower forecasting errors. This indicates that the proposed ICT-NSI is able to properly measure economic fluctuations and outperform the ICT-BSI index.

5. Conclusion and Discussion

In order to overcome the deficiencies of structured data, recent studies have increasingly used unstructured data such as text data for forecasting. In this study, we proposed a 4-step framework to construct an economic index of the ICT industry. First, under a certain criterion, we collected all news articles related to the ICT industry from 2010 to 2021. Next, we constructed a model that predicts prospect scores and sentiment scores for each sentence of the data. Then, based on the scores, we clustered the articles into six clusters and calculated cluster-weighted sentiment scores for each article. Last, by aggregating these articles' scores weekly, monthly, or quarterly, we obtain the economic index for the ICT industry.

By comparing our index and the real quarter-on-quarter GDP growth rates of the ICT industry, it is clear to see that the trend of our index shows a similar tendency to the real GDP. Also, compared to other economic indices used in South Korea, such as ICT-BSI, our index shows better forecasting accuracy. We can see that while ICT-BSI does not reflect the fluctuations of real ICT-GDP well, the proposed ICT-NSI reflects fluctuations better, especially for the results of out-sample forecasting. This indicates that our proposed method can construct an index that accurately reflects the volatility of the industry.

Since analyzing unstructured data is arduous given its complicated properties, we processed our textual news data through several rigorous steps of statistical methodologies and the application of deep learning techniques to ensure our final prediction is able to provide a benchmark to judge the overall health of an economy. Based on the experiment results by comparing our proposed method with actual ICT industry performance in South Korea, we can state that our proposed method is able to measure ICT-related economic activities by reflecting the fluctuation of the economic activity around its long-term potential level. Furthermore, the framework can also be applied to several research fields such as economic policy evaluation and consumer sentiment analysis.

Acknowledgement

J. Im is supported by the National Research Foundation (NRF) Korea, NRF-2021R1C1C1014407.

Table Appendix

Table A1

Dataset structure with variable description

Field Name	Example	Description
yearmonth	201001	Year + Month
news_id	0	Year + Unique number for each month
date	2010/1/1	Date
week	1	Number of week
media	Segye Daily	Name of media
section	IT General	Category name from Naver news
title	DDR3 DRAM, mainstream in the PC market since 2010	Title of news article
topic_id	6	Topic number of news article
sentence_seq	1	Sentence number (0: title, 1: first sentence, ...)
sentence_cnt	9	Number of sentences of news article
sentence	The existing method has a fuel shortage problem,...	Sentence of news article

Table A2

Training dataset

news_id	date	title	sentence
27	2010/1/1	LG has invested 15 trillion won this year and increase sales by 8%	The aim is to provide financial products and services of the best quality, ...
27	2010/1/1	LG has invested 15 trillion won this year and increase sales by 8%	Even in the midst of the global financial crisis, captial markets have ...
27	2010/1/1	LG has invested 15 trillion won this year and increase sales by 8%	In addition, implementation of the captial markets act has ...
...
116328	2021/12/31	IMR completes urban noise data set based on living noise analysis	Next year's AI HUB will open data for AI, ...
116328	2021/12/31	IMR completes urban noise data set based on living noise analysis	Users can use the AI program for city sound datasets, ...
116328	2021/12/31	IMR completes urban noise data set based on living noise analysis	Also, with the urban sound dataset, the noise of construction sites ...

Table A3

Relevance Ratio of Each Category

Relevance	IT	Games	General	Econ	Science	Global	Econ	Finance	Mobile	Security	Industry	Local	Econ	Internet	SME	Stock	Computer	Telecom
Irrelevant	0.474	0.429	0.925	0.860	0.905	0.960	0.334	0.476	0.840	0.945	0.529	0.869	0.894	0.445	0.358			
Relevant	0.526	0.571	0.075	0.140	0.095	0.040	0.666	0.524	0.160	0.055	0.471	0.131	0.106	0.555	0.642			

Table A4

Weight Matrix for MCDA

Category	IT	Games	General	Econ	Science	Global	Econ	Finance	Mobile	Security	Industry	Local	Econ	Internet	SME	Stock	Computer	Telecom
Weight	0.017	0.042	0.108	0.075	0.100	0.125	0.058	0.008	0.067	0.117	0.025	0.083	0.092	0.033	0.05			

Table A5

Example of Quarterly Index

Year	Quarter	Index
2010	1	127.683
2010	2	137.311
2010	3	129.792
...
2021	2	126.984
2021	3	125.125
2021	4	138.303

Table A6

Example of Monthly Index

Year	Month	Index
2010	1	130.566
2010	2	88.652
2010	3	121.987
...
2021	10	125.776
2021	11	141.400
2021	12	144.073

References

- [1] Aguilar, P., Ghirelli, C., Pacce, M., Urtasun, A., 2021. Can news help measure economic sentiment? an application in covid-19 times. *Economics Letters* 199, 109730. doi:10.1016/j.econlet.2021.109730.
- [2] Belton, V., Stewart, T.J., 2002. Multiple criteria decision analysis: an integrated approach. Kluwer Academic Publishers, Boston.
- [3] Bok, B., Caratelli, D., Giannone, D., Sbordone, A.M., Tambalotti, A., 2018. Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics* 10, 615–643. URL: <https://doi.org/10.1146/annurev-economics-080217-053214>, doi:10.1146/annurev-economics-080217-053214.
- [4] BOLIS, J.M., 2014. Eureka - eurostat review on national accounts and macroeconomic indicators. URL: <https://ec.europa.eu/eurostat/cros/content/eureka-eurostat-review-national-accounts-and-macroeconomic-indicators-en>.
- [5] Calomiris, C.W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. *Journal of Financial Economics* 133, 299–336. doi:10.1016/j.jfineco.2018.11.009.
- [6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.
- [7] Fraiberger, S.P., 2016. News sentiment and cross-country fluctuations. *SSRN Electronic Journal* doi:10.2139/ssrn.2730429.
- [8] Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics* 55, 665–676. URL: <https://www.sciencedirect.com/science/article/pii/S0304393208000652>, doi:10.1016/j.jmoneco.2008.05.010.
- [9] Hwang, C.L., Lai, Y.J., Liu, T.Y., 1993. A new approach for multiple objective decision making. *Computers Operations Research* 20, 889–899. URL: <https://linkinghub.elsevier.com/retrieve/pii/030505489390109V>, doi:10.1016/0305-0548(93)90109-V.
- [10] Jung, Y., 2021. A Study on the Social Development Index and Forecasting Model for ICT and Health and Welfare Policies. Korea Information Society Development Institute (KISDI).
- [11] of Korea, E.I.B., 2021. 2021 quarterly report of ict industry 2021-ICT. URL: <https://keri.koreaexim.go.kr/HPHFOE052M01/64059?curPage=2>.
- [12] Kuzin, V., Marcellino, M., Schumacher, C., 2011. Midas vs. mixed-frequency var: Nowcasting gdp in the euro area. *International Journal of Forecasting* 27, 529–542. URL: <https://www.sciencedirect.com/science/article/pii/S0169207010000427>, doi:10.1016/j.ijforecast.2010.02.006.
- [13] Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents. URL: <https://arxiv.org/abs/1405.4053>, doi:10.48550/ARXIV.1405.4053.
- [14] Lehmann, R., 2015. Survey-based indicators vs. hard data: What improves export forecasts in Europe? 196. URL: <https://www.econstor.eu/handle/10419/108763>.
- [15] Li, S., Li, R., Peng, V., 2021. Ensemble albert on squad 2.0 URL: <http://arxiv.org/abs/2110.09665>. arXiv:2110.09665 [cs].
- [16] Lloyd, S., 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28, 129–137. doi:10.1109/TIT.1982.1056489.
- [17] Mariano, R.S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real gdp. *Oxford Bulletin of economics and statistics* 72, 27–46. doi:10.1111/j.1468-0084.2009.00567.x.
- [18] Mehdi Keshavarz Ghorabae, Edmundas Kazimieras Zavadskas, L.O.Z.T., 2015. Multi-criteria inventory classification using a new method of evaluation based on distance from average solution (edas). *Informatica* 26, 435–451. doi:10.15388/Informatica.2015.57.
- [19] Németh, B., Molnár, A., Bozóki, S., Wijaya, K., Inotai, A., Campbell, J.D., Kaló, Z., 2019. Comparison of weighting methods used in multi-criteria decision analysis frameworks in healthcare with focus on low- and middle-income countries. *Journal of Comparative Effectiveness Research* 8, 195–204. URL: <https://www.futuremedicine.com/doi/10.2217/ce-2018-0102>, doi:10.2217/ce-2018-0102.
- [20] Pamucar, D.S., Tarle, S.P., Parezanovic, T., 2018. New hybrid multi-criteria decision-making dematel-mairca model: sustainable selection of a location for the development of multimodal logistics centre. *Economic Research-Ekonomska Istraživanja* 31, 1641–1665. URL: <https://www.tandfonline.com/doi/full/10.1080/1331677X.2018.1506706>, doi:10.1080/1331677X.2018.1506706.
- [21] Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar. pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>, doi:10.3115/v1/D14-1162.
- [22] Shapiro, Adam H., S.M., Wilson, D., 2020. Measuring news sentiment , 01–49doi:10.24148/wp2017-01.
- [23] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- [24] Stević, , Pamučar, D., Puška, A., Chatterjee, P., 2020. Sustainable supplier selection in healthcare industries using a new mcdm method: Measurement of alternatives and ranking according to compromise solution (marcos). *Computers Industrial Engineering* 140, 106231. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0360835219307004>, doi:10.1016/j.cie.2019.106231.
- [25] Szalavetz, A., 1998. The reliability of hard indicators for measuring restructuring performance. *Eastern European Economics* 36, 05–27. URL: <https://www.tandfonline.com/doi/full/10.1080/00128775.1998.11648658>, doi:10.1080/00128775.1998.11648658.
- [26] Thokala, P., Devlin, N., Marsh, K., Baltussen, R., Boysen, M., Kalo, Z., Longrenn, T., Mussen, F., Peacock, S., Watkins, J., Ijzerman, M., 2016. Multiple criteria decision analysis for health care decision making—an introduction: Report 1 of the ispor mda emerging good practices task force. *Value in Health* 19, 1–13. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1098301515051359>, doi:10.1016/j.jval.2015.12.003.
- [27] Thorndike, R.L., 1953. Who belongs in the family? *Psychometrika* 18, 267–276. URL: <http://link.springer.com/10.1007/BF02289263>, doi:10.1007/BF02289263.
- [28] Thorsrud, L.A., 2016. Nowcasting using news topics. big data versus big bank. *SSRN Electronic Journal* URL: <https://www.ssrn.com/abstract=2901450>, doi:10.2139/ssrn.2901450.
- [29] Yates, F., 1948 241, 345–377. doi:10.1098/rsta.1948.0023.
- [30] Yoon, K., 1987. A reconciliation among discrete compromise solutions. *Journal of the Operational Research Society* 38, 277–286. URL: <https://www.tandfonline.com/doi/full/10.1057/jors.1987.44>, doi:10.1057/jors.1987.44.
- [31] Yu, P.L., 1973. A class of solutions for group decision problems. *Management Science* 19, 936–946. URL: <https://pubsonline.informs.org/doi/10.1287/mnsc.19.8.936>, doi:10.1287/mnsc.19.8.936.
- [32] Zavadskas, E.K., Turskis, Z., 2010. A new additive ratio assessment (aras) method in multicriteria decision-making. *Technological and Economic Development of Economy* 16, 159–172. URL: <https://journals.vilniustech.lt/index.php/TEDE/article/view/5850>, doi:10.3846/te.2010.10.