IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 April 2025, accepted 15 May 2025, date of publication 23 May 2025, date of current version 4 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3572908

RESEARCH ARTICLE

Data-Driven Policy Making Framework Utilizing TOWS Analysis

YEONBIN SON[1], (Graduate Student Member, IEEE), YOUNGEUN KIM[2], AND YERIM CHOI[3,4]

[1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22903, USA
[2]Department of Administration, Sungkyunkwan University, Jongno-gu, Seoul 03063, South Korea
[3]Department of Data Science, Ewha Womans University, Seodaemun-gu, Seoul 03760, South Korea
[4]ai.m Inc., Seocho-gu, Seoul 06648, South Korea

Corresponding author: Yerim Choi (yrchoi@ewha.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) (Artificial Intelligence Convergence Innovation Human Resources Development, Ewha Womans University) under Grant RS-2022-00155966.

ABSTRACT In the era of hyper-uncertainty, the demand for data-driven policy making has grown significantly, as it enables the fast and accurate capture of complex phenomena. This contrasts with conventional policy making processes, which are often slow and heavily reliant on the limited experience and knowledge of a select group of experts. Although interest in data-driven approaches continues to grow, existing research primarily focuses on identifying policy agendas, leaving policymakers to develop the specific content for those agendas from scratch. In this paper, we propose a novel data-driven policy making framework based on TOWS analysis, a well-established method for strategic extraction that systematically integrates internal strengths and weaknesses with external opportunities and threats. Our framework automatically extracts policy agendas and their specific contents. It begins by identifying national capabilities and cross-cutting issues to establish key agendas using topic modeling and sentiment analysis. Then, a TOWS matrix is constructed to derive policy directions by linking capabilities with issues. Here, relevant phrases are extracted to elaborate on these directions, forming policy content. To validate our framework, we conducted a case study on South Korea's post-COVID-19 policies. The policies derived from our framework closely aligned with real-world ones and, in some instances, provided more detailed recommendations. This demonstrates the potential of data-driven approaches to enhance the precision and comprehensiveness of policy development in times of uncertainty.

INDEX TERMS Data-driven policy making, TOWS analysis, SWOT matrix, TOWS matrix, topic modeling, sentiment analysis.

I. INTRODUCTION

Recently, there has been a growing interest in data-driven policy making, an approach that designs policies based on the evidence [1]. According to Dunn [2], the policy life cycle consists of four stages as shown in Fig. 1: agenda setting, policy content construction, confirmation by major boards, and policy implementation. Policies are guidelines established by the government to achieve specific objectives aimed at addressing societal problems, and the policy life cycle encompasses the setting of these objectives and the selection of appropriate instruments to resolve issues identified in the policy agenda [3], [4]. The first two stages–agenda setting and policy content construction– are collectively referred to as the policy making process, and this involves defining national-level problems and objectives, as well as identifying the means to achieve these goals. Normally, the conventional policy making process is lengthy and lacks objectivity, since a group of experts from various sectors collaborates to derive policies based on their own experience and knowledge.

In this era of hyper-uncertainty, data-driven policy making is becoming increasingly indispensable as challenges grow more complex and unpredictable. Unexpected and

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

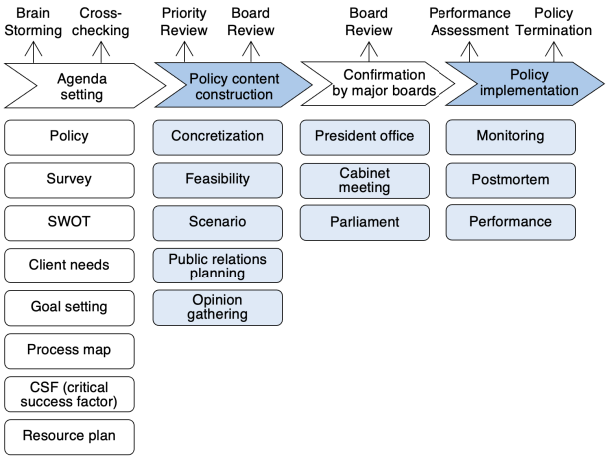


FIGURE 1. Policy life cycle. It consists of four stages: agenda setting, policy content construction, confirmation by major boards, and policy implementation.

unprecedented events, such as climate change, pandemics, accidents, terrorism, and cyberattacks, have highlighted the limitations of conventional approaches, since they heavily rely on the experts' experience and knowledge. Moreover, these approaches are often less effective when dealing with challenges that exceed prior experiences. From this perspective, it is obvious that data-driven policy making is more beneficial in quickly and accurately understanding such unexpected phenomena through data analysis and using the insights gained from that to policy making.

Several researchers have conducted studies to suggest the ideal process of data-driven policy making frameworks to identify policy issues [1], [5], [6], [7]. These frameworks generally consist of three steps. The first step is data collection and preprocessing, where policy-relevant data is gathered from various data sources, including public census data, interviews, and crowdsourced data. The second step is agenda setting, which involves identifying policy problems. Lastly, the third step is policy content construction, which focuses on determining the strategies to achieve policy goals. The last two steps are usually conducted by using data science methods.

However, existing research primarily focuses on the agenda setting stage, with limited progression toward the policy content construction, where policies are actually developed. Some studies have explored identifying key terms relevant to policy making through basic statistical analysis [8]. Others have examined representative national agendas and gauged public sentiment toward those agendas using topic modeling or sentiment analysis [9], [10]. Despite these advancements, policymakers still need to manually draft policies from scratch.

In response to these challenges, we propose a data-driven policy making framework that adopts the TOWS analysis process, a strategic tool commonly used to evaluate and align internal strengths and weaknesses with external opportunities and threats, thereby formulating actionable strategies. Traditionally, this method relies on manual analysis performed by experts from various fields. However, our goal is to automate this process using data science methods.

The framework consists of two steps: N-SWOT (national-SWOT) factor derivation and P2S (phrase-to-strategy) matrix construction. The first step aims to determine agendas necessary for policy making. To achieve this, we perform topic modeling on reports published by major policy institutes to identify national capabilities and cross-cutting issues. Then, sentiment analysis is conducted on each to learn strengths/weaknesses and opportunities/threats based on their polarity. In the second step, we use n-gram analysis to identify the content frequently discussed in reports related to each N-SWOT factor. This helps us understand which policy contents should be addressed for each agenda. Finally, a P2S matrix (our version of the TOWS matrix) is visualized with the extracted phrases, and policymakers complete the matrix by formulating strategies.

The main contributions of our study are summarized as follows. First, we propose a data-driven policy making framework that automatically identifies agendas and policy content in need of redesign. Based on this analysis, the framework provides guidance for policymakers. Next, the effectiveness of the proposed framework was validated through a case study on post-COVID-19 policy making in South Korea, and the results were compared with policies that were actually implemented.

The rest of the paper is organized as follows. Section II explains the methods necessary for understanding the proposed framework. The proposed framework is introduced and explained in Section III. Section IV presents a case study that applies the proposed framework to the post-COVID-19 context. The paper concludes in Section V.

II. BACKGROUND

A. TOWS ANALYSIS

TOWS analysis is an extended version of SWOT analysis [11] commonly used for strategic planning in organizations [12]. This approach goes beyond simply identifying SWOT factors by focusing on developing actionable strategies, making it particularly useful in scenarios where various factors are intricately interrelated. Originally, SWOT analysis was designed to help organizations identify and evaluate their strengths, weaknesses, opportunities, and threats [11]. It examines both internal characteristics (strengths and weaknesses) and external conditions (opportunities and threats) [13]. By organizing these elements into a SWOT matrix, organizations can visually prioritize factors, facilitating better strategic decision-making and planning.

Once the SWOT factors are identified, a TOWS matrix is constructed to develop strategies efficiently. This matrix consists of four quadrants, with external factors placed on the horizontal axis and internal factors on the vertical axis. Through this structure, organizations can devise strategies to leverage internal strengths to exploit external opportunities

or to mitigate external threats by addressing internal weaknesses. Specifically, S-O strategies (strengths-opportunities) aim to use strengths to capitalize on opportunities, while S-T strategies (strengths-threats) focus on using strengths to counter external threats. W-O strategies (weaknesses-opportunities) seek to improve weaknesses to seize opportunities, and W-T strategies (weaknesses-threats) aim to minimize weaknesses to reduce or avoid threats.

Several studies have utilized TOWS analysis to derive policies or corporate strategies. For instance, TOWS analysis was employed in [14] to design Canadian energy and climate policy. Another study applied TOWS analysis to explore appropriate policies for the post-COVID-19 era [15]. Other studies have combined data science methods with TOWS analysis. One such study extracted frequently occurring phrases (bi-grams and tri-grams) from online customer reviews of smart devices and used these as a basis for experts to perform TOWS analysis [16]. In [17], the authors performed topic modeling on online reviews about a university to identify key topics of student concern, and then experts executed TOWS analysis based on topic modeling results to discuss the university's development direction.

B. DATA SCIENCE METHODS

1) TOPIC MODELING

Topic modeling is a statistical technique employed to uncover abstract topics within a collection of documents [18]. Its primary goal is to automatically identify patterns of word usage and link documents that exhibit similar patterns [19]. By extracting latent variables from large datasets, topic modeling helps discern trends in the usage of words and phrases [20]. This method enables the classification of each document into multiple topics. It is especially valuable for analyzing large datasets, offering insights that would be challenging to obtain through manual analysis [21].

This technique is widely applied across various fields. In marketing, it generates recommendations and collects user opinions [22], [23], [24]. It aids in organizing databases of documents into groups based on their similarity [20]. Additionally, it is used to compare user interactions on Twitter with those on other media platforms and to analyze user-generated content on LEGO ideas [25]. Furthermore, it helps evaluate online student reviews to identify the main topics discussed about the university [17].

Topic modeling methods are broadly divided into two groups. The first group includes NMF (non-negative matrix factorization) methods [26]. These methods learn topics by decomposing the word-document matrix into two low-rank factor matrices: a document-topic matrix and a topic-word matrix. NMF performs well for high-dimensional data tasks such as dimension reduction and clustering [26], [27]. However, the topics generated by NMF can be challenging to interpret due to their representation.

The second group consists of generative probabilistic models, including LSA (latent semantic analysis) [28], PLSA (probabilistic latent semantic analysis) [29], and LDA (latent Dirichlet allocation) [30]. These models discover latent semantic structures from a text corpus without requiring any prior annotations or labeling of the documents. In these models, each document is assumed to be a mixture of topics, with a distribution over words characterized by each topic. Generative probabilistic models have significantly impacted organizing text groups for articles, research papers, and user-generated text such as reviews. However, they showed poor performance on short texts [31].

2) SENTIMENT ANALYSIS

Sentiment analysis is a NLP (natural language processing) technique used to identify and extract people's opinions, emotions, or attitudes toward an entity from text data [32], [33]. It typically classifies text into positive, neutral, or negative sentiments. Automatically analyzing user-generated data is crucial for monitoring public opinion and aiding decision-making [34]. This technique gauges public sentiment by analyzing textual data, such as product reviews [35], social media posts [36], [37], news articles [38], or political documents [39].

The process of sentiment analysis can be conducted at various levels depending on the target: document-level, sentence-level, phrase-level, and aspect-level. Document-level analysis evaluates the sentiment of an entire document to determine an overall polarity. Sentence-level analysis breaks down a document into individual sentences and determines the polarity of each, which is useful for documents containing mixed sentiments [40]. Phrase-level analysis is beneficial for text with multiple lines, such as product reviews. Aspect-level analysis focuses on all aspects within a sentence, determining the polarity for each aspect.

Applications of sentiment analysis are diverse and span multiple fields. In marketing, it helps improve products and services by analyzing customer feedback [33]. According to [5], news articles can be examined to gauge public opinion, aiding in the formulation of government strategy. In the finance sector, [41] utilized sentiment analysis to understand market trends and make informed investment decisions.

Several methods are used to perform sentiment analysis. The lexicon-based approach involves assigning predefined scores to each lexicon according to its polarity, such as 1 for positive, 0 for neutral, and -1 for negative [35]. By scoring each lexicon (token) in a sentence and calculating the final score, the overall polarity is assigned to the text. The machine learning approach utilizes shallow and deep learning methods such as Naive Bayes, SVM (support vector machine), RNN (recurrent neural network), and LSTM (long short-term memory) for developing a sentiment classifier. In supervised learning, models are trained on labeled data where polarity is the label, and these models predict sentiment. Recently, LLMs (large language models) have been applied to sentiment analysis, demonstrating superior performance. Transformer-based models such as BERT [42]

and GPT-3 [43] are utilized to represent text, and the representations are used to train classifiers.

III. DATA-DRIVEN POLICY MAKING FRAMEWORK BASED ON TOWS ANALYSIS

A. OVERVIEW

We propose a data-driven policy making framework by adopting the TOWS analysis process, which is frequently used in conventional policy making, based on data science methods. The first step is to derive N-SWOT factors from reports published by policy institutes, using topic modeling and sentiment analysis. Previous studies have shown that when performing topic modeling, setting a relatively small number of topics yields abstract topics, while setting a larger number of topics generates detailed topics [21], [44], [45]. Since this step aims to find the policy agenda, we perform topic modeling twice with different parameters: a small number for national capabilities and a large number for cross-cutting issues. Then, through sentiment analysis, we classify national capabilities as strengths/weaknesses and cross-cutting issues as opportunities/threats.

The second step involves constructing our newly designed TOWS matrix, the P2S matrix. An example is shown in Fig. 2 (b). For policymakers, it is essential to review the key issues being discussed in relation to agendas that require policy decisions. To assist with this, we perform n-gram analysis to explore the most frequently mentioned phrases associated with each N-SWOT factor, and present them in the P2S matrix. Policymakers use this matrix to examine the main discussion points within a specific agenda and subsequently complete the matrix by formulating strategies.

	Positive	Negative
National capabilities	Strengths	Weaknesses
Cross-cutting issues	Opportunities	Threats

(a)

National capabilities / Cross-cutting issues	Strengths	Weaknesses
	(Topic 1) Phrase 1, Phrase 2, ... (Topic 2) Phrase 1, Phrase 2, ... (Topic N) Phrase 1, Phrase 2, ...	(Topic 1) Phrase 1, Phrase 2, ... (Topic 2) Phrase 1, Phrase 2, ... (Topic N) Phrase 1, Phrase 2, ...
Opportunities	Strength & Opportunity	Weakness & Opportunity
(Topic 1) Phrase 1, Phrase 2, ... (Topic 2) Phrase 1, Phrase 2, ... (Topic N) Phrase 1, Phrase 2, ...	What strengths can government use to capitalize on country's opportunities?	What weaknesses must be mitigated to capitalize on country's opportunities?
Threats	Strength & Threat	Weakness & Threat
(Topic 1) Phrase 1, Phrase 2, ... (Topic 2) Phrase 1, Phrase 2, ... (Topic N) Phrase 1, Phrase 2, ...	What strength can government use to better handle country's threats?	What weaknesses can be used by country's external threats?

(b)

FIGURE 2. Example of (a) N-SWOT matrix and (b) P2S matrix.

B. DATA COLLECTION & PREPROCESSING

Reports published by policy institutes are used to eliminate any possibility of noise and bias. These reports solely present factual information and exclude the writer's interests or stereotypes. As a result, readers can gain a clear understanding of the current situation and the issues at hand. Furthermore, as these documents aim to propose strategies for the future based on current circumstances, analyzing them helps identify topics discussed during the most recent period.

For data collection, if institutions provide reports directly on bulletin boards on their websites, we web-crawl the text, and if they provide the reports as files, we download them. After acquiring all the uploaded reports, files that do not contain the text subject to analysis, such as images or Excel files for statistics, are filtered out. If the report is provided as a PDF, we use OCR (optical character recognition) technology [46] to convert files into editable and searchable data.

Next, we perform morpheme extraction as a part of data preprocessing. A morpheme is the smallest unit of meaning [47]. We tokenize the document and tag each token with Part-Of-Speech (POS). Only meaningful tokens, such as nouns, adjectives, and adverbs, are retained.

Since general words that do not reflect the topic of the report become noise in the analysis, we conduct stopwords removal. This process involves constructing a stopwords dictionary and removing the words the dictionary contains. Additionally, for more effective removal, we utilize a domain-specific stopwords dictionary in addition to widely used open-source dictionaries. For constructing the domain-specific dictionary, words are sorted by word frequency or document frequency and add the top few words to the dictionary. Here, the number of words is decided by the analyst. The final report set obtained after data-preprocessing is denoted as R, and $R = \{r_1, \cdots, r_k, \cdots, r_K\}$, where r_k and K indicate $k - th$ report and the number of reports.

C. N-SWOT FACTORS DERIVATION

Now, we analyze the preprocessed report set R to derive N-SWOT factors. As we described in Section III-A, two rounds of topic modeling with different parameters are executed to extract national capabilities and cross-cutting issues, following the TOWS analysis process, which involves identifying an organization's internal and external factors and formulating strategies based on them. The final N-SWOT factors are then learned by performing sentiment classification to assess each factor's polarity, as shown in Fig. 3.

After topic modeling, we obtain two topic sets. One is the set of national capabilities by $C = \{c_1, \cdots, c_m, \cdots, c_M\}$, with the relatively small number of topics as a parameter, and the other is the set of cross-cutting issue is defined as $I = \{i_1, \cdots, i_l, \cdots, i_L\}$, with the relatively large number of topics as a parameter. Here, c_m and M indicate $m-$th national capability and the number of topics, while i_l and L respectively represent $l-$th cross-cutting issue topic and the number of topics. Typically, the optimal number of topics is determined using quantitative metrics such as coherence scores or perplexity. However, in our

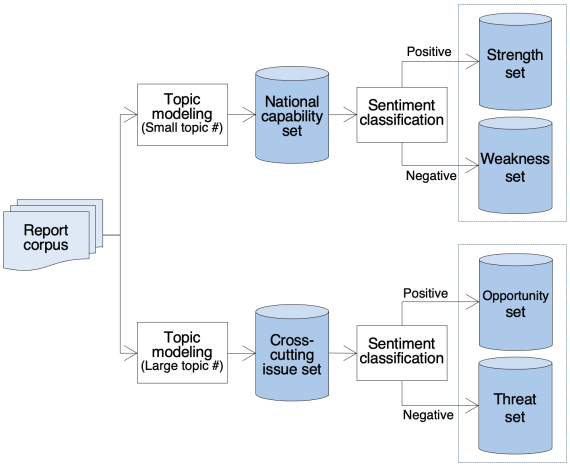


FIGURE 3. Specific process for learning N-SWOT factors.

framework, it is important to extract two distinct topic sets: one capturing broader capabilities and the other for granular issues. Therefore, instead of relying on quantitative performance, we qualitatively assess the interpretability and relevance of topic sets across different topic numbers and select a representative setting.

We utilized LDA for topic modeling. LDA is a generative probabilistic model of a corpus R. Essentially, LDA assumes that there are hidden structures within the data, aiming to uncover these structures by identifying patterns in the word distributions across the reports. It allows us to identify the underlying matter that national reports mainly discuss in the recent period. In this model, each report is represented as a mixture of latent topics, and a distribution over words characterizes each topic.

LDA assumes the following process for each report r in the corpus R. First, choose $\theta \sim Dir(\alpha)$, where θ represents the topic mixture for the document. For each word h_g in r, choose a topic $z_g \sim Multinomial(\theta)$, where z_g is the topic of $g-$th word. Then, choose a word h_g from $p(h_g|z_g, \beta)$, a multinomial probability conditioned on the topic z_g. Here, α and β are corpus-level parameters, θ is a report-level variable representing the topic mixture, z_g is a word-level variable indicating the topic of the $g-$th word, and h_g is the $g-$th word itself. Gibbs sampling is employed to infer the latent topic distribution of each report and the word distribution of each topic.

Now, we perform sentiment classification on C and I. Topics within the former set that have a higher positive score than negative score are categorized into strength set S, while those with a higher negative score are categorized into weakness set W. For the latter set, topics with a higher positive score are categorized into opportunity set O, and those with a higher negative score are categorized into threat set T.

To perform classification, we construct representative documents for each topic in the topic sets. For each national capability topic $(c_1, \cdots, c_M)$, we calculate the cosine similarity value between a topic and a report. As a result, we obtain $U = [u_{mk}]_{M\times K}$, where u_{mk} indicates a similarity value between $m-$th national capability topic and $k-$th report. Then, for making $m-$th topic's representative document, if $k-$th column of U's maximum value is u_{mk}, r_k would be added to p_m, which is $m-$th national capability topic's representative document. After iterating through all elements, we obtain a representative document set for the national capabilities $D_C = \{d_{c_1}, \cdots, d_{c_m}, \cdots, d_{c_M}\}$, where d_{c_m} represents the document associated with the $m-$th capability. A similar process is applied to the set of cross-cutting issues F_I to derive the corresponding representative document set $D_I = \{d_{i_1}, \cdots, d_{i_l}, \cdots, d_{i_L}\}$, where d_{i_l} indicates to the document for the $l-$th cross-cutting issue.

Now, we perform sentiment classification for each representative document using LLM. Each document in D_C and D_I (d_{c_m} and d_{i_l}) is input into the LLM for generating a high-dimensional vector embedding. The embeddings capture the semantic nuances and contextual information of the text, however, its quality depends on the dataset on which the LLM was trained. Therefore, in our framework, since the dataset used is the set of policy reports, we manually label sentences and use them to fine-tune the pre-trained LLM. This process helps the model better understand policy-related terminology and context, enabling it to learn specific sentiment tags more accurately. Once the embeddings are obtained, they are fed into a sentiment classifier, such as shallow or deep learning models, which have been trained to distinguish between polarities (positive, neutral, and negative). The output is the polarity of each representative document.

D. P2S MATRIX DEVELOPMENT AND POLICY MAKING

The goal of this step is to develop a P2S matrix and use it to make policy making. Unlike the traditional TOWS matrix, which only listed SWOT factors and relies on experts to determine the direction for deriving strategies, our matrix provides phrases frequently associated with each national factor. These phrases help to directly deduce strategies for policy making. With these information, policymakers can gain a more detailed understanding of the policy content they need to focus on when designing policies.

To obtain the phrases that are frequently mentioned in reports for each factor, n-gram analysis is performed. n-gram analysis is a computational linguistic technique that involves breaking down text into contiguous sequences of n words in length [48]. For example, the bigrams of the sentence "I love Seoul" are "I love" and "love Seoul". This technique is effective in understanding context and structure more accurately by splitting a large text into short contiguous sequences. By combining all representative documents of topics corresponding to each factor into the factor's representative document, and then performing n-gram analysis on that document, only the top phrases based on frequency are retained.

The P2S matrix includes four policy strategies: S-O strategy, S-T strategy, W-O strategy, and W-T strategy. The S-O strategy combines the topics identified in the strengths of national resources with the opportunity of cross-cutting issues. For example, if ''high technology (artificial intelligence, research and development, start-up)'' represents an opportunity in cross-cutting issue, and ''education (learning support, youth, university)'' is a strength of national resources, the S-O strategy could be ''strengthening technological innovation by organizing the curriculum to focus on high technology, such as artificial intelligence.'' The S-T strategy combines the strengths of national resources with the threats of cross-cutting issues. For instance, if the threat is ''immigration (migrant workers, multiculturalism, foreign employment),'' the S-T strategy might involve ''strengthening domestic integration by providing learning support to foreign and multicultural families.''

The W-O strategy aligns the weaknesses of national resources with the opportunities presented by cross-cutting issues. For example, if the weakness in national resources is the ''economy (online shopping, user protection, vulnerable groups),'' the W-O strategy could involve ''developing financial services tailored to vulnerable groups using new technologies.'' The W-T strategy integrates the weaknesses of national resources with the threats of cross-cutting issues. In this case, the strategy could be ''reforming consumer protection laws for financially vulnerable groups, such as foreigners.''

As demonstrated in the example above, the P2S matrix serves as a critical role for structuring policy content. The N-SWOT matrix, which includes an assessment of national capabilities, resources, and the environment, is used to set the policy agenda. The P2S matrix enables decision makers to prioritize objectives and refine specific tasks for policy making. In conclusion, the P2S matrix provides a framework for both the government and policy experts to identify agendas, set objectives, and prioritize key policy issues necessary for the formulation of the national strategic plan, the nation's highest-level policy document.

IV. CASE STUDY

To validate the effectiveness of our proposed framework, we conducted a case study focused on policy making in the post-COVID-19 era. The analysis was based on reports published by major policy institutes in South Korea.

A. DATA COLLECTION & PREPROCESSING

For the experiments, we utilized reports published by a total of 44 Korean major policy institutes from January 20, 2020 to January 31, 2021. To ensure accurate analysis for policy making in the post-COVID-19 era, we focused on approximately one year during which COVID-19 emerged and reached its peak. Reports from 26 institutes were collected from NKIS (national knowledge information system) [49], a digital service portal for searching and utilizing research outputs from government-funded research institutes. Reports from 7 institutes were gathered from EIEC (economic information and education center) [50], which collects and analyzes domestic and international economic information to support rational economic activities of the public. The remaining 11 institutes' reports were obtained directly from their websites. When reports were provided in PDF format, Google OCR technology [51] was used. The process involved converting each PDF file page into PNG format and extracting the text from the PNG images.

After securing the dataset, we preprocessed it by removing stop words and filtering documents. Three dictionaries were employed for stopwords removal: two open source dictionaries [52], [53] and one domain-specific dictionary. Frequency analysis was performed on tokens extracted through morphological analysis to build the domain-specific dictionary. Words within the top 3,000 by word frequency and the top 1,000 by document frequency were added to the dictionary. Since this case study was specialized for the post-COVID-19 era, only COVID-19 related reports were used. Reports containing the relevant terms 'COVID-19, corona virus disease 19, Corona Virus, Post Corona, Post COVID-19, Pandemic, Endemic, Epidemic' were retained. A total of 9,843 documents were gathered, with 6,148 of them being post-COVID-19 related.

B. N-SWOT FACTORS DERIVATION

First, two parameters are needed to be determined for topic modeling: the number of words to be analyzed and the number of topics. The first parameter was set to 33,750. This is the number of all unique words that appeared less than 15 times across all documents. For the number of topics, we tested a range of topics from 5 to 40 and ultimately selected a relatively small number of 7 and a larger number of 35. As described in Section III-C, these values were chosen based on a qualitative evaluation of topic interpretability and relevance, rather than relying on quantitative metrics.

For sentiment classification, a fine-tuned KoBERT model was utilized. We got representative documents' representation based on KoBERT [54] and used them to determine the polarity of each sentence in a document by comparing the positive score to the negative score. KoBERT was developed to overcome BERT's performance limitation on Korean text [42], and it was pre-trained primarily on the Korean Wikipedia corpus. While KoBERT achieved a benchmark accuracy of 92%, we observed a significant performance drop to 82% accuracy when applying the model to policy-related reports. To address this, we fine-tuned the model using a domain-specific dataset of 8,376 sentences from public reports.

We conducted experiments to validate the use of fine-tuned KoBERT. The test set consisted of 400 randomly selected sentences from reports, with 200 labeled as positive and 200 labeled as negative by experts. We compared the performance against sentiment classifiers from the Kakao NLP framework and KoBERT's sentiment classifier. The

TABLE 1. Topic modeling results: Each topic name was labeled by policy experts after analyzing the assigned words within each topic.

Topic set	Topics
Set of national capabilities	ICT, Infrastructure, Job, Education, Welfare, Trade, Diplomacy
Set of cross-cutting issues	Politics, Policy, Law, Employment, Labor, Welfare, Maritime, Education, Trade, Community, Transportation, Finance, Urban Planning, Manufacturing, Student Rights, Curriculum, Crime, Agriculture, Youth, Advanced Technology, Energy, ICT, COVID-19, Media, Elderly Welfare, Child Welfare, Finance, Conflict, Inter-Korean Relations, Household, Housing Policy, Immigration, Childcare, Marine, Environmental Pollution

performance evaluation metric was the F1 score. On policy institutes' reports, KoBERT pre-trained on the Korean Wiki dataset achieved an F1 score of 81%, while Kakao's classifier, pre-trained on an e-commerce review corpus, scored 78%. In contrast, our fine-tuned KoBERT model achieved an F1 score of 91 %, representing an improvement of up to 14%.

Now, we present topic modeling results based on South Korea's reports related to COVID-19. The national capability set C and the cross-cutting issue set I are shown in Table 1. C and I contain 7 and 35 topics, respectively. In the table, we can find topic names that were labeled by policy experts based on the word set for each topic.

Then, we got the final N-SWOT factors by performing sentiment analysis on C and I as shown in Fig. 4. In terms of C, the strengths (S) include the advances in ICT, education, and infrastructure (transport/mobility, energy, skills/talent, regulatory improvements, etc.) The weaknesses (W) involve challenges in trade and diplomacy, such as trade disputes, geopolitical risks, and cultural tourism issues. Social issues, such as aging populations, unemployment, social polarization, and gaps in support for vulnerable groups (the elderly, women, multicultural families, and children), are also highlighted. Furthermore, employment challenges related to economic and industrial restructuring, income inequality, and gaps in vocational training are emphasized.

In terms of I, the opportunities (O) lie in leveraging ICT for emerging technologies, advancing digital education, and fostering mobility innovations, including autonomous driving, hydrogen/electric vehicles, and supporting economic and financial recovery. Other opportunities include revitalizing media content and platforms, promoting participatory governance through public-private partnerships, regulatory reforms, enhancing regional competitiveness and community regeneration, investing in R&D for emerging technologies, and accelerating the transition to a green economy, particularly with regard to ecosystem and ocean sustainability. Threats (T) include the declining competitiveness of small businesses due to the rise of online shopping and the shift to contactless transactions. Additionally, the widening education gap, strained north-south relations, job insecurity, elderly poverty, urban aging, agricultural decline, weakening manufacturing competitiveness, housing insecurity, youth challenges, and the adverse effects of pollution and climate change are identified as significant threats.

C. P2S MATRIX DEVELOPMENT AND POLICY MAKING

In this step, South Korea's N-SWOT factors were used to construct the P2S matrix as shown in Fig. 5. We extracted frequently mentioned phrases for each topic within the reports, which policymakers used as references to establish strategies. For instance, to capitalize on the nation's strengths and seize opportunities, leveraging ICT tools, such as AI and digital platforms, is crucial for expanding educational opportunities for vulnerable groups. New jobs should be created through ICT-driven economies, smart infrastructure, and technical education. Furthermore, digital transformation and regulatory reform can strengthen both trade and foreign policy, while welfare improvements can be achieved by upgrading infrastructure and providing retraining opportunities for older adults. Data-driven policies are crucial for enhancing the efficiency of policy-making and addressing social issues related to welfare and employment.

To address South Korea's weaknesses, it is necessary to boost digital literacy and ensure education is accessible to vulnerable groups by utilizing online resources. Given the significant job market challenges, job creation should be facilitated through the development of smart infrastructure and vocational training in high-tech sectors. Trade and diplomatic relations must be improved through regulatory reform and the convergence of new technologies. Moreover, now is the time to revitalize communities and improve the quality of life through local development and environmental conservation, while addressing social issues through media content and data-driven solutions.

For handling South Korea's threats with strength points, efforts should also focus on supporting vulnerable populations by closing educational gaps through digital education, which helps reduce inequality. Economic recovery can be promoted by investing in smart infrastructure and vocational training programs. Moreover, it is necessary to enhance trade and diplomatic relations through digital transformation and adherence to international standards. The social safety net can be expanded by addressing welfare policy gaps through targeted education and infrastructure improvements.

Finally, strengthening the social safety net requires addressing gaps in welfare coverage, particularly for vulnerable groups, by introducing digital finance solutions Educational disparities can be reduced by offering online education and customized programs designed for those

TABLE 2. Comparison of the New Deal 2.0 action plan implemented during the post-COVID-19 era and the strategy derived from our proposed framework.

	Subject	New Deal 2.0 Action Plan	Our Strategy
Digital	ICT innovation & digital transformation	* Strengthening the D.N.A ecosystem Build, develop, and utilize data related to citizens' lives and various sectors Expand the convergence of 1st, 2nd, and 3rd industrial revolutions with 5G and AI Create AI-based smart government Build a K-AI cybersecurity system	* Building data-driven policies to promote the digital economy Leverage data from ICT and digital transformation to gain insights for policy improvement and enhance policy efficiency Develop and refine data-driven policies to tackle social issues in welfare and employment, providing practical solutions to these challenges
	Smart infrastructure & industrial competitiveness	* Advancing digital infrastructure Build key digital infrastructure in early childhood education and lifelong learning Build smart medical and remote healthcare infrastructure Support online businesses for small and medium enterprises (SMEs)	* Enhancing technological innovation and support measures Provide retraining opportunities for older adults through lifelong learning programs and media education Promote technological development and innovation to counteract the decline in manufacturing competitiveness, agricultural reduction, and pursue diversification of related industries Accelerate the adoption of smart factories to support sustainable agricultural technologies
	Diplomatic relation & geopolitical risk	None	* Improving trade and diplomatic risks through regulatory reform Use digital platforms to respond effectively to global supply chain crises and stabilize diplomatic relations through swift measures Simplify trade processes and ensure compliance with international standards using technology solutions to mitigate trade disputes and geopolitical tensions
Green	Green innovation & green economy	* Building a low-carbon economy foundation Establish a monitoring system for carbon emissions in manufacturing and expert industries Expand carbon-neutral systems and related infrastructure for key sectors Promote awareness of carbon neutrality among the public and support the expansion of eco-friendly practices	* Regulatory improvement and urban regeneration to address environmental pollution and climate change Improve aging infrastructure and resolve urban decay issues through smart city technologies, advanced transportation policies, and energy innovations Expand urban regeneration projects and establish ecosystem-centered urban planning to resolve urban decay and housing instability Improve illegal street vending and licensing systems to ensure pleasant urban spaces
	Decentralized energy & sustainable planning	* Expanding low-carbon, decentralized energy Build smart grids for efficient energy use Expand the use of new and renewable energy and convert public energy sources * Promoting green innovation and building eco-friendly ecosystems Expand public support for green and eco-friendly industries Strengthen research, development (R&D), and financial foundations for green industries	* Addressing environmental challenges through infrastructure improvement Tackle pollution and climate change through green infrastructure solutions and sustainable urban planning Build eco-friendly infrastructure and improve regulations to address environmental pollution and climate change Develop eco-friendly solutions to address rural environmental issues such as livestock manure management
Human	Human investment & social safety net stability	* Human investment Strengthen training for human resources in software and AI Train skilled professionals for future industries Strengthen education systems for agriculture and rural industries through digital transformation * Employment and social safety net stability Strengthen employment safety net and social insurance for various job sectors Enhance employment support for vulnerable groups Support workers impacted by unemployment and career transitions	* Developing policies to strengthen communities and improve prosperity Address aging and social issues by enhancing local competitiveness and revitalizing communities Improve the quality of life and social inclusion for local residents through local development policies and conservation of the natural environment
	Bridging the gap & supporting vulnerable groups	* Youth policy Expand future-oriented jobs and entrepreneurship opportunities for young people Establishing a foundation for youth livelihood stability * Bridging the gap Strengthen educational support for low-income and underprivileged youth Enhance healthcare and support systems for the vulnerable Increase cultural participation opportunities for citizens	* Reducing educational gaps and supporting vulnerable groups Provide economic and social support to vulnerable groups (such as elderly poverty and youth unemployment) by utilizing artificial intelligence, the digital economy, and smart technologies Expand and enhance educational opportunities for vulnerable groups by leveraging ICT tools such as artificial intelligence, digital transformation, edtech, and cyber education, as well as through targeted student support Enhance support for rural areas and small-scale classes through online education platforms to improve educational disparities and learning environments

most in need. Employment stability can be supported by diversifying vocational training and introducing work-hour reduction policies. Manufacturing and agricultural competitiveness can be enhanced through technological innovation and the adoption of smart factories. Moreover, improving regulations and infrastructure is crucial for addressing environmental pollution and mitigating climate change challenges.

Now, we compared the strategies derived from our framework with the actual policies implemented during the same period to assess the feasibility of the proposed approach. The comparison of the policies deducted from our framework and the actual policy is present in Table 2. We evaluated whether the proposed policy was suitable for the post-COVID-19 era. The subject of the evaluation is the Korean New Deal, a large-scale national initiative aimed at fostering

	Positive	Negative
National capabilities	(ICT) The ICT field is actively investing in research and development, focusing on data platforms and data literacy. (Education) As educational facilities are turned to online class due to COVID-19, interest in new education policies is increasing. (Infrastructure) Plans for securing human resources and developing future-oriented infrastructure are newly emerging.	(Welfare) Women's economic activities have been restricted, and the elderly poverty rate has increased. (Job) Due to the economy crisis, companies are dismissing employees, and the issue of insufficient job training facilities has emerged. (Diplomacy) Cultural and tourism exchanges were halted, and relations between major countries deteriorated as they focused on national efficiency, stability, and resilience due to the pandemic. (Trade) The trade industry has experienced the most significant economic slowdown due to COVID-19, making it urgent to establish a system to address this issue.
Cross-cutting issues	(Curriculum) A suitable curriculum and a system for online classes are necessary for students. (Transportation) A new transportation and toll-cuts are emerged to promote the local economy. (Economy) The overall economy is recovering through monetary policies. (Child welfare) Childcare policy is revised due to the culture of respect for parenting. (Energy system) Energy transition to prevent climate change is accelerated. (Housing policy) The government is establishing policies to expand infrastructure across various sectors. (Community) Various plans for supporting local community is developed. (Marine) There is a high demand for smart logistics in the era of digitization. (Social media) Online platform is growing fast because of people's lifestyle change. (Advanced technology) There is vibrant investment in research and development for future-oriented science and technology.	(Household) Trends have changed due to pandemics, making it urgent to reform policies that guarantee social equity. (Inter-Korean relations) The relations are deteriorated due to the denuclearizing issue. (Immigration) Social conflict is increasing due to discrimination against immigrants and the neglect of hatred. (Elderly welfare) There is a lack of system to maintain a quality of life above the minimum level. (COVID-19) The effectiveness of implemented policies is being evaluated. (Urban planning) There is a limitation of regeneration due to the urban aging. (Trade) Poor progress results from unexpected situations in NPS countries. (Crime) There is a lack of improvement in the human rights protection policy for crime victims. (Environmental pollution) Preventing the emergence of climate change caused by environmental pollution is essential. (Employment) High qualifications are required to find work due to a lack of job opportunities. (Manufacturing) The new trend is emerged in manufacturing industry. (Agriculture) The rural economy has declined as agricultural workers' incomes have decreased. (Labor) Employment and wage inequality caused by wage decrease. (Student rights) It is necessary to guarantee the youth's life and rights.

FIGURE 4. N-SWOT matrix by analyzing post-COVID-19 related reports published by South Korea's major policy institutes.

innovative growth in the post-COVID-19 era. This policy was promoted by the Korean government for about a year following the COVID-19 outbreak. The evaluation focuses on the Korean New Deal 2.0 action plan, announced in July 2021, which was released after the data collection period of this study (January 2020 to January 2021).

Based on the comparison, the policy agenda in the P2S matrix largely aligned with the actual policies. Notably, key agendas emphasized in the actual policies—such as new technology convergence, digitization of social infrastructure, promotion of green industries (e.g., environment and energy), strengthening employment and social safety nets, and addressing social disparities and inequality—were aligned with the results from our framework. Additionally, the inclusion of policies for regionally balanced growth, which were absent from the initial Korean New Deal, appeared to reflect keywords such as ''rural crisis'' and ''community recovery,'' which our study identified as threats and opportunities in cross-cutting issues. However, there are some differences between our strategies and the actual policies. Our framework emphasizes the need for strategies to address political diplomacy and geopolitical risks, such as foreign policy regarding North-South Korea relations and trade policy concerning the U.S.-China trade conflict. It also

Cross-cutting issues \ National capabilities	Strengths	Weaknesses
	(Infrastructure) Skilled workforce, transportation policy, urban regeneration, smart city, regulatory reform (Education) Education policy, media education, edu-tech, digital twin, research and development investment (ICT) Platform development, data utilization, digital transformation	(Jobs) Vocational education and training, skill development, worker employment, working hours decrease (Diplomacy) Agricultural agreements, cultural tourism, diplomatic relations (Trade) Trade disputes, North Korea sanctions, industrial regulation, institutional arrangements (Welfare) Elderly poverty, women's economic participation, social problem-solving, multicultural family children
Opportunities	**Strength & Opportunity**	**Weakness & Opportunity**
(Marine) Green economy, infrastructure sustainability, marine mammal protection (Community) Planning management, nature conservation, regional development (Media) digital platform, data economy, social impact, literacy education curriculum (Energy) Energy transition, climate change, renewable energy, solar & wind power (Technology convergence) AI application, intelligent government, digital economy (Governance) Proactive administration, regulatory reform, citizen participation (Infrastructure policy) Public institution collaboration, job measures, urban regeneraton (Child welfare) Private education monitoring, parenting support, childcare services (Transportation) Toll reduction, autonomous vehicles, mobility service innovation (Marine logistics) Fishing vessel safety, smart ports, maritime big data (Advanced technology) Technology development, startup support, R&D investment	• Strengthening support for vulnerable groups through ICT and education • Job creation through smart infrastructure and the digital economy • Addressing trade and foreign policy issues through regulatory reform and digital transformation • Improving welfare through education and infrastructure • Building data-driven policies to promote the digital economy	• Improving digital literacy and well-being of vulnerable groups • Creating jobs through smart infrastructure and economic recovery • Improving trade and diplomacy through regulatory reform and technological convergence • Developing policies to strengthen communities and improve prosperity • Solving social problems using media content and data
Threats	**Strength & Threat**	**Weakness & Threat**
(Crime) Video recording, victim protection, crime prevention activities (Learning environment) Career development, vulnerable groups, academic achievement (Healthcare welfare) Telemedicine, mental health, retirement income (Agriculture) Livestock waste management, farm income decline (Inclusive finance) Digital finance, consumer protection, tax policy (Urban planning) Urban regeneration, ecosystem development, permit system (Housing policy) Public rental housing, housing rights protection, vulnerable groups (Inter-Korean relations) Inter-Korean agreements, exchange & cooperation projects (Household) Small businesses, online shopping, ready-made meal production (Manufacturing) Technological development, minimum wage, labor demand reduction (Environmental pollution) Microplastics, climate change response, waste disposal (Student rights) Youth rights, academic anxiety, youth support (Elderly welfare) Social security system, aging society, elderly poverty (COVID-19) Disease prevention, social distancing, emergency disaster relief funds (Labor) Working hours reduction, women's employment, career interruption (Trade agreement) New Southern Policy, international cooperation, national economy (Employment) Human resource development, skill development, institutional support (Online education) Curriculum, learning opportunities, academic underachievement (Immigration) Multicultural education, migrant workers, visa status (Parenting policy) Parenting support, workplace childcare, infant care (Gender equality) Gender equality, workplace sexual harassment, women's status	• Mitigating social risks by supporting vulnerable groups through digital education • Addressing urban aging and environmental challenges through infrastructure improvement • Promoting economic recovery through smart infrastructure and technological innovation • Improving trade and diplomatic risks through regulatory reform • Addressing social inequality through inclusive economic policies	• Strengthening the social safety net by addressing welfare blind spots for vulnerable groups • Reducing educational gaps and supporting vulnerable groups through online education • Institutional support to address labor market instability and ensure employment stability • Enhancing manufacturing and agricultural competitiveness through technological innovation and support measures • Regulatory improvement and urban regeneration to address environmental pollution and climate change

FIGURE 5. P2S matrix for policy-making in Korea's post-COVID-19 era. Strategies were established based on the N-SWOT factors and the frequently mentioned phrases associated with each factor.

highlights the need for strategies to revive tourism, including measures to address the decline in foreign visitors post-COVID-19. In contrast, the actual policies do not include provisions for foreign policy between South and North Korea, trade policy in response to U.S.-China trade conflicts and geopolitical risks, or tourism policy.

V. CONCLUSION

We propose a novel data-driven policy making framework that enables policymakers to swiftly and accurately assess unexpected phenomena and provide clear direction for formulating policy in an era of hyperuncertainty. Unlike existing data-driven approaches, which focus on analyzing data to identify key national agendas—leaving policymakers to develop policy from scratch—our framework goes a step further by automatically developing a P2S matrix. This matrix refines policy content in relation to the main agenda, offering clear guidance on which areas and which content to prioritise. The value of our framework is enhanced by the use of reliable data sources, particularly reports published by policy institutes.

We also conducted a case study using reports from South Korea's major policy institutes, specifically published during the post-COVID-19 era. Upon comparing the strategies derived from our framework with the policies that were actually implemented (New Deal 2.0 Action Plan), we found numerous similarities. Moreover, in the area of diplomatic relations, we observed that our framework proposed a more advanced policy.

Our future research directions are as follows: while the current framework focuses on deriving broad policy directions, the next step—identifying specific policy issues—will require collaboration with policy experts. Therefore, the goal of future research is to ensure that a data-driven framework guides all stages of the policy making process. In addition, we plan to enhance the current content identification process, which is based on n-gram analysis. As an extension of this study, we will incorporate recent transformer-based text summarization methods to extract more refined phrases frequently associated with each agenda. Finally, while the present study validated the proposed framework through a case study using South Korea's policy documents, future work will apply ours to English-language datasets to assess its applicability in diverse policy environments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] M. A. Hossin, J. Du, L. Mu, and I. O. Asante, "Big data-driven public policy decisions: Transformation toward smart governance," *Sage Open*, vol. 13, no. 4, Oct. 2023, Art. no. 21582440231215123.

[2] W. N. Dunn, *Public Policy Analysis*. New York, NY, USA: Routledge, Jul. 2015.

[3] A. Hanberger, "What is the policy problem: Methodological challenges in policy evaluation," *Eur. Eval. Soc.*, vol. 7, no. 1, pp. 45–62, Jan. 2021.

[4] W. K. Foundation., *Logic Model Development Guide*, vol. 7. Battle Creek, MI, USA: W.K. Kellogg Foundation, 2000.

[5] A. F. van Veenstra and B. Kotterink, "Data-driven policy making: The policy lab approach," in *Electronic Participation*. Cham, Switzerland: Springer, 2017, pp. 100–111.

[6] F. Mureddu, J. Schmeling, and E. Kanellou, "Research challenges for the use of big data in policy-making," *Transforming Government: People, Process Policy*, vol. 14, no. 4, pp. 593–604, Apr. 2020.

[7] A. Papazoglou Chalikias, I. Tsampoulatidis, F. Tsalakanidou, S. Nikolopoulos, I. Kompatsiaris, N. Komninos, K. Doudouliakis, G. Papastergios, P. Papafilis, S. Karkaletsi, and C. Chatzis, "Evidence-driven policy-making using heterogeneous data sources—The case of a controlled parking system in Thessaloniki," *Data Policy*, vol. 2, p. e15, Jan. 2020.

[8] Y. Son, S. Kim, and Y. Choi, "ICT trend analysis based on research papers and patents," *J. Korea Converg. Soc.*, vol. 12, no. 12, pp. 1–18, Dec. 2021.

[9] J. R. Grande-Ramírez, A. A. Aguilar-Lasserre, G. A. Arrioja-Carrera, and J. E. Domínguez-Herrera, "Applications and resources for social media sentiment analysis: A strategic planning case study," in *Management and Industrial Engineering Economies*. Cham, Switzerland: Springer, 2024, pp. 121–156.

[10] J. Lee, E. Im, I. Yeo, and G. Gim, *A Study on the Strategy of SWOT Extraction in the Metavers Platform Review Data: Using NLP Techniques*. Cham, Switzerland: Springer, 2023, pp. 173–188.

[11] T. Hill and R. Westbrook, "SWOT analysis: It's time for a product recall," *Long Range Planning*, vol. 30, no. 1, pp. 46–52, Feb. 1997.

[12] H. Weihrich, "The TOWS matrix—A tool for situational analysis," *Long Range Planning*, vol. 15, no. 2, pp. 54–66, Apr. 1982.

[13] K. R. Andrews and D. K. David, *The Concept of Corporate Strategy*, vol. 3. Homewood, IL, USA: Irwin, 1987.

[14] C. Fertel, O. Bahn, K. Vaillancourt, and J.-P. Waaub, "Canadian energy and climate policies: A SWOT analysis in search of federal/provincial coherence," *Energy Policy*, vol. 63, pp. 1139–1150, Dec. 2013.

[15] H. Jiang, "A SWOT analysis of smart governance applications amid the COVID-19 pandemic," *Inf. Polity*, vol. 28, no. 3, pp. 377–393, Sep. 2023.

[16] S. Rajendran and E. Pagel, "Insights on next-generation manufacturing of smart devices using text analytics," *Heliyon*, vol. 6, no. 7, Jul. 2020, Art. no. e04491.

[17] S. Srinivas and S. Rajendran, "Topic-based knowledge mining of online Student reviews for strategic planning in universities," *Comput. Ind. Eng.*, vol. 128, pp. 974–984, Feb. 2019.

[18] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, Dec. 2020, Art. no. 101582.

[19] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 1, pp. 147–153, 2015.

[20] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook Latent Semantic Analysis*. New York, NY, USA: Psychology Press, 2007, pp. 439–460.

[21] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1445–1456.

[22] Y. Son, H. An, and Y. Choi, "A proofreader matching method based on topic modeling using the importance of documents," *J. Internet Comput. Services*, vol. 19, no. 4, pp. 27–33, Aug. 2018.

[23] M. Rossetti, F. Stella, and M. Zanker, "Analyzing user reviews in tourism with topic models," *Inf. Technol. Tourism*, vol. 16, no. 1, pp. 5–21, Mar. 2016.

[24] P. Kumar, A. K. Kushwaha, A. K. Kar, Y. K. Dwivedi, and N. P. Rana, "Managing buyer experience in a buyer–supplier relationship in MSMEs and SMEs," *Ann. Oper. Res.*, Sep. 2022, doi: 10.1007/s10479-022-04954-3.

[25] D. Schlagwein and N. Bjorn-Andersen, "Organizational learning with crowdsourcing: The revelatory case of LEGO," *J. Assoc. Inf. Syst.*, vol. 15, no. 11, pp. 754–778, Nov. 2014.

[26] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[27] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proc. World Wide Web Conf. World Wide Web (WWW)*. Lyon, France: ACM Press, 2018, pp. 1105–1114.

[28] S. Dumais, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, Jan. 2004.

[29] T. Hofmann, "Probabilistic latent semantic indexing,," in *Proc. 22nd Annual Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Berkeley, CA, USA, Aug. 1999, pp. 50–57.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[31] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, Mar. 2022.

[32] V. Dhanalakshmi, D. Bino, and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," in *Proc. 3rd MEC Int. Conf. Big Data Smart City (ICBDSC)*, Piscataway, NJ, USA, Mar. 2016, pp. 1–5.

[33] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Lang. Process. J.*, vol. 2, Mar. 2023, Art. no. 100003.

[34] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[35] Y. Son, Y. Roh, S. Kim, and Y. Choi, "A matrix factorization based manuscript editor recommendation method considering ratings and sentiment score of clients," in *Korean Institute of Information Scientists and Engineers*. Jeju Island, South Korea: Korean Institute of Information Scientists and Engineers, 2019.

[36] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marqués, "Data-driven strategies in operation management: Mining user-generated content in Twitter," *Ann. Oper. Res.*, vol. 333, nos. 2–3, pp. 849–869, Feb. 2024, doi: 10.1007/s10479-022-04776-3.

[37] U. Shahzad, T. Sengupta, A. Rao, and G. D. Sharma, "Energy in the backseat? Investigating decarbonization dialogue in supply chain tweets during and after COVID-19," *Ann. Oper. Res.*, Jan. 2024, doi: 10.1007/s10479-023-05806-4.

[38] T. Xu, Q. Peng, and Y. Cheng, "Identifying the semantic orientation of terms using S-HAL for sentiment analysis," *Knowl.-Based Syst.*, vol. 35, pp. 279–289, Nov. 2012.

[39] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decis. Support Syst.*, vol. 53, no. 4, pp. 680–688, Nov. 2012.

[40] B. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2014, pp. 325–335.

[41] M.-Y. Day and C.-C. Lee, "Deep learning for financial sentiment analysis on finance news providers," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 1127–1134.

[42] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jan. 2018, pp. 2–18.

[43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jan. 2020, pp. 1877–1901.

[44] M. Samejima, "Topic analysis of case reports in tourism towards collaborative tourism planning support," in *Tourism Informatics: Towards Novel Knowledge Based Approaches*. Berlin, Germany: Springer, 2015, pp. 1–14.

[45] A. Bittermann and J. Rieger, "Finding scientific topics in continuously growing text corpora," in *Proc. 3rd Workshop Scholarly Document Process.* Gyeongju, South Korea: Association for Computational Linguistics, Oct. 2022, pp. 7–18.

[46] G. Nagy, T. A. Nartker, and S. V. Rice, "Optical character recognition: An illustrated guide to the frontier," *Proc. SPIE*, vol. 3967, pp. 58–69, Dec. 1999.

[47] C. Anderson, *Essentials Linguistics*. Hamilton, ON, Canada: McMaster Univ., Mar. 2018.

[48] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Develop.*, vol. 1, no. 4, pp. 309–317, Oct. 1957.

[49] N. R. Council. *National Knowledge Information System*. Accessed: Apr. 5, 2025. [Online]. Available: https://www.nkis.re.kr/eng/main.do
[50] K. D. Institute. *Economic Information and Education Center*. Accessed: Apr. 5, 2025. [Online]. Available: https://www.kdi.re.kr/eng/introduce/eiec
[51] *Google Cloud AI*. Accessed: Apr. 5, 2025. [Online]. Available: https://cloud.google.com/use-cases/ocr
[52] *Stopwords-Ko*. Accessed: Apr. 5, 2025. [Online]. Available: https://github.com/stopwords-iso/stopwords-ko
[53] *Korean Stopwords*. Accessed: Apr. 5, 2025. [Online]. Available: https://www.ranks.nl/stopwords/korean
[54] W. Kim, B. Son, and I. Kim. (2020). *KoBERT: Pretrained Language Models for Korean*. [Online]. Available: https://github.com/SKTBrain/KoBERT



YEONBIN SON (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in industrial and management engineering from Kyonggi University, Suwon, South Korea, in 2018 and 2020. She is currently pursuing the Ph.D. degree with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. Her research interests include machine learning-based recommender systems and the use of human performance modeling.



YOUNGEUN KIM received the B.S. degree in statistics and economics finance from Keimyung University, South Korea, in 2017, and the M.S. degree in statistics from Kyungpook National University, South Korea, in 2019. She is currently pursuing the Ph.D. degree in public administration with the Graduate School of Governance, Sungkyunkwan University, South Korea. From 2019 to 2025, she was a Senior Researcher with the Department of Artificial Intelligence Policy, the National Information Society Agency (NIA), South Korea. Her research interests include machine learning, public governance, and AI policy.



YERIM CHOI received the B.S. and Ph.D. degrees in industrial engineering from Seoul National University (SNU), South Korea, in 2010 and 2016, respectively. She is currently an Associate Professor with the Department of Data Science, Ewha Womans University, South Korea. Her research interests include human modeling, smart manufacturing, AI applications, and deep learning.