IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 November 2023, accepted 26 November 2023, date of publication 30 November 2023,
date of current version 5 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337808

RESEARCH ARTICLE

Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews

YUSUNG AN[1], DONGJU KIM[1], JUYEON LEE[1], HAYOUNG OH[2], JOO-SIK LEE[2], AND DONGHWA JEONG[3]

[1]Sungkyunkwan University, Seoul 03063, South Korea
[2]College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, South Korea
[3]MechaSolution, Daegu 41024, South Korea

Corresponding author: Hayoung Oh (hyoh79@gmail.com)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant NRF-2022R1F1A1074696; in part by the Technology Innovation Program (or Industrial Strategic Technology Development Program-Source Technology Development and Commercialization of Digital Therapeutics) (Development of Digital Therapeutics for Depression from COVID19) funded by the Ministry of Trade, Industry and Energy (MOTIE), South Korea, under Grant 20014967; in part by the BK21 FOUR Project (Bigdata Research and Education Group for Enhancing Social Connectedness Thorough Advanced Data Technology and Interaction Science Research) funded by the Ministry of Education (MOE), South Korea, under Grant 5199990913845; in part by NRF; in part by the AI Convergence Research Fund; in part by Sungkyunkwan University, 2023; in part by the SMC-SKKU Future Convergence Research Program Grant and the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) under Grant RS-2023-00254129; and in part by the Graduate School of Metaverse Convergence, Sungkyunkwan University.

ABSTRACT Reviews left by consumers on e-commerce platforms provide crucial marketing information as they are a publicly available source of information providing insight into consumers' thoughts and opinions. However, it is physically impossible to read hundreds of reviews per product. Therefore, algorithmically extracting the necessary insights from a large volume of review data can provide more advanced information while reducing time and costs. This study aims to automate the process of extracting related products, pros and cons of products, and trend forecasting from review data using clustering algorithms. The review dataset for pros and cons of products and related products was constructed by selecting 17 products on the Naver Shopping platform and crawling them and the product keyword and search volume dataset for trend forecasting was constructed by crawling the data from Itemscout and Naver datalab platform. Various clustering-based algorithms such as Deep Clustering Network, BERTopic, and a Transformer-based forecasting model were used to conduct the research. It is expected that this will allow for a more accurate understanding of consumer thoughts, which can be utilized in marketing for various products and services.

INDEX TERMS Topic modeling, clustering algorithm, trend forecasting, BERT, e-commerce marketing.

I. INTRODUCTION

In the era of digital technology, individuals post and distribute reviews on diverse products and services via online platforms. This reservoir of review information has evolved into a valuable resource for businesses, offering insights into consumer sentiments and inclinations, which can subsequently enhance their marketing strategies. Nevertheless, effectively analyzing substantial quantities of review data and extracting meaningful insights from them presents a formidable challenge.

This research proposes a framework for extracting and analyzing marketing-related information based on review data using clustering algorithms and a Transformer-based forecasting model as a core technique. We attempted to extract marketing insights from review data using DCN and BERTopic [1] in our previous study. Here, we aim to advance the content of the study and propose a framework for analyzing various forms of information based on review data by incorporating additional features. Clustering algorithms are

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

used widely in marketing, such as for customer segmentation and recommendation systems, as they group data with similar characteristics. This process involves utilizing user log data, such as purchased products and visited pages. However, these data are not publicly available and are dependent on e-commerce platforms and sales companies, making it practically impossible for third parties to access this data.

In this study, we introduce a methodology that leverages clustering algorithms applied to external sources' review data to identify consumer intent and unveil the interconnections among products. Through the clustering of consumer-provided review data, we can group similar consumer sentiments, thereby facilitating the identification of consumer purchase intent and the parallels and relevance between products. We also introduce a trend forecasting technique based on search volume data and Transformer to predict the people's interest and demand about some products in the future. These give companies crucial insights into the relationships among specific products that can be used to develop effective sales strategies. Moreover, we outline a procedure for distilling product advantages and drawbacks from review data. Employing topic modeling, we categorize review data into specific topics and extract the merits and demerits associated with each topic. This allows companies to pinpoint the aspects of products that receive the most favorable evaluations from consumers, as well as areas that need improvement. This information can serve as invaluable reference material for product development and the formulation of marketing strategies. Through this approach, companies can more accurately understand consumer opinions and preferences, enabling them to utilize it for product and service improvements as well as the development of marketing strategies.

II. RELATED TECHNIQUES

A. CLUSTERING ALGORITHM

Clustering algorithms are used widely in various fields. One widely known clustering algorithm is the K-means algorithm. The K-means algorithm begins by choosing k cluster centroids and then iteratively adjusts their positions to minimize the following cost function:

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

where k is the number of clusters, S is the set of clusters, $S_1, S_2, \ldots, S_k$, and μ_i is the centroid of S_i [2]. This study will also utilize this algorithm.

Another clustering algorithm to be used in this study is HDBSCAN [3]. HDBSCAN algorithm from a range of hierarchical clustering algorithms. This algorithm is a variation of the DBSCAN algorithm [4] within the hierarchical framework. It calculates the mutual reachability for all pairs of data points and forms a minimum spanning tree using this data, creating a hierarchical structure for clustering. This method excels in recognizing clusters with various shapes, surpassing the capabilities of the K-means algorithm.

An essential factor to consider when conducting clustering analysis is the dimensionality of the data. If the data has a high dimensionality, achieving satisfactory clustering outcomes can be challenging due to the curse of dimensionality. Consequently, in scenarios where data dimensionality is excessively high, it is customary to integrate dimensionality reduction algorithms such as UMAP into the clustering process.

Another technique for reducing data dimensionality involves the application of autoencoders to generate latent vectors from the data. The original approach designed for this purpose is referred to as Deep Embedded Clustering (DEC) [5]. DEC employs stacked autoencoders to derive lower-dimensional latent vectors from the input data, which are subsequently utilized for clustering, exhibiting commendable performance. Deep Clustering Network (DCN) enhances DEC by adjusting the loss function, thereby achieving slightly improved performance [6]. This approach also utilizes autoencoders for dimensionality reduction and leverages them for clustering.

The innovation in this research is the introduction of a novel loss function created by amalgamating the loss functions of Autoencoders and the K-means algorithm. This fresh loss function is applied during the dimensionality reduction phase to produce vectors that are more amenable to the K-means algorithm, resulting in enhanced clustering performance. The formula for the novel loss function used in this study is provided below.

$$\min \sum_{i=1}^{N} \left(l\left(g\left(f\left(x_i\right)\right), x_i\right) + \frac{\lambda}{2} ||f\left(x_i\right) - Ms_i||_2^2 \right)$$

The component on the left enclosed by the sigma symbol signifies the autoencoder's loss function, which quantifies the disparity between the input data and the reconstructed output data. In our study, we employed Mean Square Error (MSE) for this purpose, although alternative functions like L1-norm or KL-divergence are viable options. On the right-hand side is the structure of the loss function for the K-means algorithm. By introducing a coefficient, denoted as lambda, its magnitude can be adjusted to exert varying levels of influence and control.

B. TOPIC EXTRACTION

Topic modeling represents a natural language processing method applied to unveil concealed themes within textual information. Textual data encompasses diverse subjects, and each document might have connections to multiple themes. The process of topic modeling autonomously detects these themes and gauges the extent to which each document comprises them. Among the array of topic modeling techniques, Latent Dirichlet Allocation (LDA) [7] is the most widely adopted method. LDA operates on the assumptions that documents consist of multiple topics, each document comprises a blend of these topics, and topics are characterized by distributions of words. It iteratively deduces the distribution of topics within documents and the distribution of words within topics.

While LDA performs reliably and is used widely, it has limitations in its ability to capture semantic relationships among words.

To address these limitations, one topic modeling approach that has emerged is BERTopic [8]. As the name implies, BERTopic harnesses the power of BERT. It begins by embedding textual data into vectors using Sentence BERT [9] and subsequently conducts clustering on these vectors to conduct topic modeling. The clustering process involves dimensionality reduction of vectors using UMAP, as previously mentioned, and the application of HDBSCAN. BERTopic offers improved performance by capitalizing on BERT for vector embedding, which considers the semantics and contextual information of the text. BERTopic is a topic modeling technique that employs BERT embeddings and reveals that the task can be effectively handled through a clustering-based approach.

In this study, in addition to HDBSCAN, which serves as the default clustering algorithm in BERTopic, other high-performing clustering algorithms, such as DCN, have been explored as alternatives to produce results and assess their performance. By integrating various clustering algorithms into the core framework of BERTopic, this research aims to investigate and contrast the outcomes obtained. This endeavor will furnish insights into the efficacy and performance of diverse clustering algorithms within the domain of topic modeling.

Dynamic topic modeling, a variant of topic modeling, is a method that divides data into units of a certain period of time and extracts topics for each period so that you can see changes in topics over time. In the context of LDA-based dynamic topic modeling, data is partitioned according to predefined time periods, and topics are extracted for each period based on the data using methods similar to LDA [10]. However, in this process, the word distribution utilized incorporates parameters from the preceding time period's topic modeling, ensuring that the topics at the current time point align with those from the previous time point. BERTopic also offers a version of dynamic topic modeling. In the dynamic topic modeling approach of BERTopic, clustering is initially performed on the entire dataset, assigning each data point to clusters representing different topics. Subsequently, the data is divided into period units, and topic keywords are extracted for each timestamp.

Text summarization is one of the ways that can be used to extract key content of a text along with topic modeling. summarization is the task of producing a summary based on text. Summarization involves the task of creating concise summaries from text, and it can be approached in various ways. In this study, we employ Abstract summarization, which generates fresh sentences by leveraging token-level relationships. This method utilizes advanced natural language generation models like BART, T5, GPT-3, and PEGASUS, with our study focusing on KoBART. KoBART, a model pre-trained on Korean news data, is based on the BART architecture, combining bidirectional and autoregressive transformers in a seq2seq structure. It was initially trained to reconstruct text corrupted by a noise function, drawing inspiration from BERT's bidirectional encoder and GPT's autoregressive decoder. BART has consistently demonstrated SOTA performance in tasks such as question answering, dialogue, and summarization.

C. TRANSFORMER-BASED TIME SERIES FORECASTING

Transformer is a revolutionary deep learning architecture introduced in the field of natural language processing (NLP). It has gained immense popularity due to its ability to handle sequential data efficiently, enabling significant advancements in machine translation, text generation, and various other NLP tasks. However, in recent times, Transformers have begun to be researched not only in the field of NLP but also in other domains such as computer vision, and they have started to demonstrate excellent performance in various fields. In the field of time series prediction, models utilizing Transformers have also emerged and have begun to exhibit superior performance compared to models introduced thus far. In a study by [11], a model named 'Informer' was proposed, based on the Transformer architecture, to address the challenges of long-term time-series forecasting. This model successfully mitigated the issues of significantly high time complexity and memory usage when using Transformers for extended forecasting by introducing novel mechanisms such as ProbSparse self-attention, self-attention distilling, and a generative-style decoder to drastically reduce the computational complexity of attention and the inference process.

In contrast, other research [12] has aimed to resolve the issues with Transformers in time-series prediction by introducing a lighter attention mechanism. Instead of addressing these problems, they replaced the attention layers with an auto-correlation mechanism that leveraged linear relationships between time-series values spaced at regular intervals. Furthermore, they improved performance by decomposing time series data into two components: one reflecting trends in the data and the other capturing seasonality.

Research [13] attempted to create time series data in the form of output from text data using Transformer, rather than directly handling time series data with Transformer. For stock price prediction, they used company annual reports as data and processed them with various Transformer-based models such as BERT, LongBERT, and utilized various regression algorithms based on them. This research suggests that not only past time series data but also text information related to it can improve prediction performance.

Study [14] attempted to predict product sales data. This study injected product information extracted from additional data about products into the decoder of basic transformer. However, this study addressed the problem of predicting the sales volume of new fashion items. They employed a new dataset named VISUALLE, which not only included sales data but also incorporated time-series data from Google Trends related to the products and product images.

This additional information is embedded to vectors and used to input of decoder. This attempt could improve the performance of forecasting.

III. RELATED RESEARCH

Topic modeling has already been used widely in various fields, especially in marketing. In one study [15], product description data were collected from an e-commerce platform that sells mobile phones, and topic modeling was conducted using LDA. The study then extracted the concept of products and the relationships between each feature using keywords extracted from each topic, constructing a knowledge graph to identify the semantic knowledge of the products. In another study [16], LDA was combined with inverse regression to propose a supervised learning-based topic modeling method and demonstrated good performance in predicting product categories using a benchmark dataset collected from an e-commerce platform. A study [17] collected user reviews from an e-commerce platform and predicted user ratings using topic modeling with LDA and sentiment analysis, achieving good performance. In another study [18], the collected reviews were subjected to topic modeling and sentiment analysis using the UTSJ model, which combines LDA with a sentiment analysis stage. Based on this, the study performed a task of identifying deceptive reviews.

Another study [19] proposed a model to detect fake reviews using Word2Vec-based LDA and sentiment analysis, based on the Amazon review dataset. Marketing insights related to relevant products were extracted using the gravity model and regression analysis have been used to extract marketing insights related to relevant products, based on cell phone transaction data from China's e-commerce platform Taobao [20]. Impressive results have been achieved by collecting user comments on logistic services and applying a text mining approach using CNN [21]. In the research [22], quantitative correlation analysis was conducted concerning urban commercial facilities based on consumer reviews.

LDA also has been utilized for topic modeling of consumer-related topics in various domains such as food distribution, food tourism, apartment interior design, and airline services in Korea, with corresponding analysis of consumer opinions [23], [24], [25], [26].

Study [27] proposed a model using an encoder-decoder structure with RNNs to predict the sales volume of new fashion items. An intriguing aspect of this study was the incorporation of multi-modal data, including product images and text data such as product descriptions, embedded into vectors and fed into the decoder. This approach provided additional information about the products, leading to improved performance compared to models trained without multi-modal data.

IV. METHODOLOGY

In this section, using real consumer review data gathered from an e-commerce platform, we will elaborate on the methodology for (1) extracting the strengths and weaknesses of individual products, (2) identifying the connections among different products, and (3) predicting product trends. For the first and second tasks, we will employ clustering algorithms and topic modeling techniques such as BERTopic and DCN. Additionally, we will assess their performance by making comparisons with conventional approaches like LDA and K-means algorithms. The final task also employs topic modeling as an auxiliary tool. We will utilize a Transformer-based forecasting model, as used in the previous research [14], to assess its performance by incorporating reviews as input data for the decoder and utilizing product features extracted through topic modeling.

A. DATA COLLECTION AND PREPROCESSING

We obtained consumer reviews from the 'Naver Shopping' platform, one of Korea's largest e-commerce platforms. We selected 17 product keywords based on their sales volume and then crawled the reviews and ratings for each product, resulting in a total of 530,877 reviews. This data was obtained in accordance with the platform's robots.txt policies. Figure 1 depicts the distribution of star ratings for the collected reviews, with the majority of them being awarded five stars. As star ratings decrease, the number of reviews also diminishes. Due to the substantial variations in review counts across star ratings, we utilized a logarithmic scale on the y-axis for better visualization.

The acquired data underwent preprocessing using the Korean corpus-trained morphological analyzer, Mecab [28], which disassembled the text into morphemes. Based on our observations, we determined that essential product-related information predominantly resides within nouns, verbs, and adjectives in the review text. Consequently, we retained only these three parts of speech (POS) as morphemes while eliminating all others. Furthermore, we excluded various Korean stopwords. Figure 1 displays the frequencies of the top ten most common POS categories derived from the morphological analysis of all the reviews, with nouns, verbs, adjectives, and adverbs being the most prominent. This observation solidifies our decision to exclusively utilize these four POS categories during the data preprocessing phase of our study.

The foundational time-series data for trend forecasting task were obtained using the Naver Data Lab's keyword search volume API. Through this API, weekly search volume trends for each product keyword were collected for the years 2019 to 2022, spanning a period of four years.

The product keywords used for collecting search volume were obtained by crawling through the product keywords provided in the ''Item Discovery'' category of 'Itemscout,' a service that offers e-commerce platform analysis. Approximately 50,000 product keywords were collected using this method. However, for some product keywords, search volume statistics were available only for a very limited period. Therefore, a subset of 33,495 product keywords was constructed by selecting those with search volume statistics available for over 100 weeks out of the total 209-week search volume collection period. Furthermore, product reviews and product

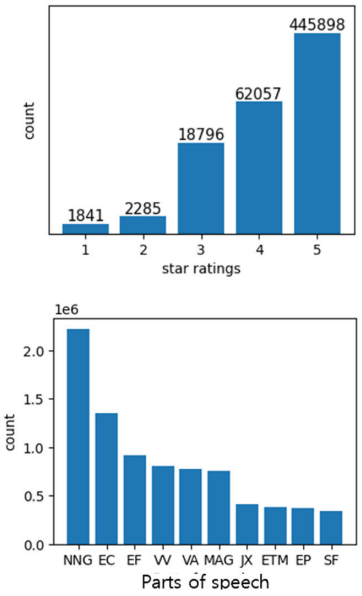


FIGURE 1. Basic information about reviews dataset. Above: Distribution of collected review data star ratings (1-5). Below: Frequency of top 10 frequently appearing parts of speech. (NNG: general noun, EC/EF/ETM/EP: ending, VV: verb, VA: adjective, MAG: adverb, JX: auxiliary participle, SF: sentence-finishing punctuation marks.)

names corresponding to the product keywords were collected by web crawling from 'Naver Shopping.'

B. EXTRACTING ADVANTAGES AND DISADVANTAGES

To extract both the positive and negative aspects, we categorized the review data based on their star ratings. Specifically, we isolated reviews rated with five stars to identify advantages and reviews with one or two stars to pinpoint disadvantages. Subsequently, we employed topic modeling techniques with these segregated review datasets.

The subsequent step involves transforming the review data into vectors. For this task, we used KcBERT [29], a BERT model trained on comments gathered from Korean online news. KcBERT exhibits strong performance in handling natural language processing tasks that involve informal text, which often includes colloquial language, slang, and typographical errors. We utilized mean pooling with KcBERT to convert the review data into vectors.

Following the embedding process, we proceeded with the task of topic modeling. Our hypothesis posits that the keywords extracted from these topics will serve as effective representatives of the product's attributes, encompassing both its strengths and weaknesses. We plan to explore the BERTopic methodology as well as a new approach that incorporates the DCN module as an alternative to BERTopic's clustering module. The process is outlined comprehensively in Fig. 2.

C. ANALYSIS OF RELATED PRODUCTS

The process of analyzing related products closely parallels the one described for extracting advantages and disadvantages. The primary distinction lies in the fact that,

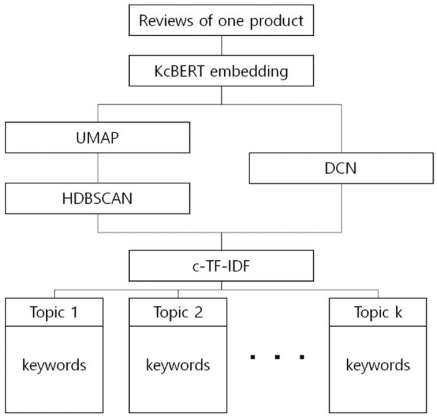


FIGURE 2. Process of extracting features of a product from review. Keywords of each topic will be features of the product.

while extracting pros and cons, we conducted clustering-based topic modeling on individual product reviews. In contrast, in the analysis of related products, we perform topic modeling using reviews from multiple products. We then examine which product reviews tend to appear together within each topic. If two products frequently co-occur in the same topic, it suggests that they share certain characteristics, indicating a relationship between the two products.

Therefore, when extracting product attributes, we prioritize assessing the co-occurrence frequency of the products using Term Frequency (TF) alone, rather than employing c-TF-IDF to identify products from clustered topics. The overall process is illustrated in Figure 3.

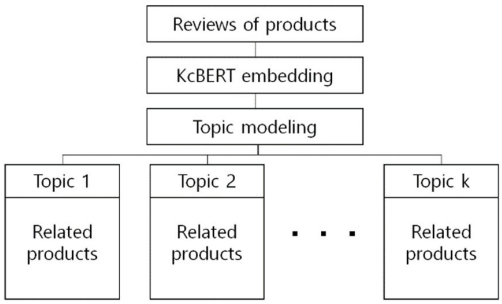


FIGURE 3. Process of extracting related products. Products co-located in the same topic will have some relations.

D. TREND FORECASTING

In the context of sales forecasting, the model employed is fundamentally based on a modification of the Transformer architecture, which was adapted from the structure proposed in a previous study for sales prediction [14]. For the initial input to the encoder, we use search volume time series data spanning 157 weeks from 2019 to 2021, associated with product keywords and their respective categories. The decoder's input consists of product keywords, reviews related to the keywords, and text extracted from the product names that represent the product features. To achieve this, we used two approach, topic modeling and summarization. First, we conduct topic modeling on the review data and

calculate TF-IDF for keywords extracted from each topic, ultimately selecting the top five keywords. In this study, BERTopic is employed for this purpose. However, in some cases, there may not be a sufficient amount of review data available for certain products. In such cases, we resort to using TF-IDF based on keywords extracted from the product names collected alongside. The extracted keywords, along with the product category information, are concatenated into a single string. Second, we conducted text summarization on review data. Thus, we were able to get one summarized sentence describing the product feature per product. Subsequently, concatenated keywords and summarization sentences are embedded into vectors using a pre-trained KcBERT [29] and passed through a basic MLP.

Finally, for the decoder's output, the subsequent sales volume trend is predicted. In this study, we configure the encoder's input to predict the estimated search volume figures for the following year, spanning 52 weeks. The loss functions utilized during the training process consist of MSE and correlation coefficient. While MSE is typically the primary choice for loss functions, for this specific task, understanding the overall trend is more practically significant than precisely predicting numerical values. In other words, in addition to minimizing the error between actual and predicted data, a strong linear relationship between the two datasets implies a better reflection of the trend in the target time series data [30]. Therefore, we define and employ a novel loss function that combines MSE and the correlation coefficient. The formulation of the loss function is as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{y}_i - y_i\right)^2 + \lambda\left(1 - Cor\left(\hat{y}_i, y_i\right)\right)\right)$$

Cor(x, y) represents the Pearson correlation coefficient between two vectors, x and y. It is a weight constant that determines the influence of the correlation coefficient values, and in this study, it was set to 200 for utilization. The overall structure of the framework, including all detailed tasks, can be found in Appendix.

V. EXPERIMENTS AND RESULTS

In this section, we will examine the results of extracting product advantages, disadvantages, and related products through the described process using real data. We will compare the performance of BERTopic and DCN, which were used in this study, with traditional methods such as LDA and K-means algorithms using various metrics. By analyzing the extracted information and evaluating the performance of different algorithms, we can gain insights into the strengths and weaknesses of each method in terms of accurately identifying product characteristics and discovering related products.

All models were trained using the optimized number of topfics determined through the hyperparameter optimization process mentioned earlier. For DCN, the training process involved two stages: pretraining using stacked autoencoders and actual DCN training. Both stages were trained for 50 epochs. In the original research, the authors encountered a problem of the trivial solution in the setting using the MNIST benchmark dataset, where all data was assigned to a single label. To address this issue, they reduced the size of the hidden layers in the autoencoder. Instead of using (500-500-2000) parameters, they used (192-192-768) hidden layer sizes, which matched the dimensions of their dataset. The dimension of the generated latent vectors was set to 10. All experiments were conducted using the NVIDIA Tesla T4 GPU on Google Colaboratory.

A. EXTRACTING FEATURES

Model performance is evaluated using three metrics. Firstly, Topic Coherence is measured using Normalized Pointwise Mutual Information (NPMI) [31]. NPMI assesses the frequency of co-occurrence between two words within a text, measured on a scale from −1 to 1. Higher values on this scale signify enhanced coherence. Topic Diversity (TD) serves as a metric for evaluating the degree of dissimilarity between the identified topics in their semantic significance [32]. This metric quantifies the extent to which distinct and semantically meaningful keywords have been extracted across various topics. Its computation involves determining the proportion of unique words across all topics, and it falls within the range of 0 to 1. Values approaching 1 signify a greater diversity among the topics. Word Embedding Coherence (WE) is an evaluation metric that involves embedding the extracted keywords as vectors and calculating the similarity between words [33]. All cosine similarities between vectors of words constituting each topic are calculated and combined. In addition, this calculation is performed for each topic, and the average of the sum of cosine similarities by topic is used. The high WE coherence value shows that the semantic similarity between the words constituting each topic is high. Therefore, it can be determined that the higher this value, the higher the Topic Coherence. The performance results of the topic modeling are presented in Table 1 and Fig. 4.

TABLE 1. Topic modeling evaluation (for one product).

	NPMI	TD	WE
LDA	**-0.017**	0.633	5.335
	-0.158	0.566	6.164
KcBERT+Kmeans	-0.026	0.366	9.662
	-0.172	0.540	**9.655**
KcBERT+DCN	-0.060	0.575	**10.211**
	-0.115	0.600	9.339
KcBERT+UMAP +HDBSCAN (BERTopic)	-0.052	**0.733**	**10.295**
	-0.303	**0.820**	8.534

The advantage extraction performance was better than the disadvantage extraction performance. This discrepancy is attributed to the dissimilarity in dataset sizes. Typically, in consumer-generated reviews, there are far fewer reviews with low ratings than high ratings. Consequently, the dataset used for extracting disadvantages was considerably smaller in size, and this limitation is evident in the results.

TABLE 2. Extracted features with topic modeling.

pros	cons
곰팡이[mold], 제거[removal], 배송[delivery], 사용[use], 효과[effectiveness], 빠르[fast], 제품[product], 청소[cleaning], 냄새[odor], 뿌리[spray]	효과[effectiveness], 생각[thought], 그래요[be], 곰팡이[mold], 분무[spray], 제거[removal], 심하[intense], 냄새[odor], 최악[worst], 빠르[fast]
빠른[fast], 배송[delivery], 청소[cleaning], 제품[product], 화장실[toilet], 사용[use], 만족[satisfaction], 욕실[bathroom], 상품[product], 곰팡이[mold]	심해요[intense], 냄새[odor], 지워짐[erased], 지워져요[erased], 심해서[intense], 실패[fail], 개인[personal], 샀어요[bought], 어지러워[dizzy], 아파요[painful]
대비[compared], 가격[price], 곰팡이[mold], 스프레이[spray], 분사[spray], 효과[effectiveness], 벽지[wallpaper], 혁명[revolution], 청소[cleaning], 할인[discount]	제거[removal], 욕실[bathroom], 가능[possible], 받자[receive], 스티커[sticker], 생긴[formed], 비누[soup], 싱크대[sink], 만남[encounter], 물총[water gun]
-	락스[bleach], 냄새[odor], 곰팡이[mold], 제거[removal], 사용[use], 벽지[wallpaper], 화장실[bathroom], 발송[delivery], 청소[cleaning], 뿌리[spray]
-	사용[use], 곰팡이[mold], 제품[product], 뿌리[spray], 바르[cover], 효과[effectiveness], 지워[erased], 똑같[same], 힘들[difficult], 구입[purchase]

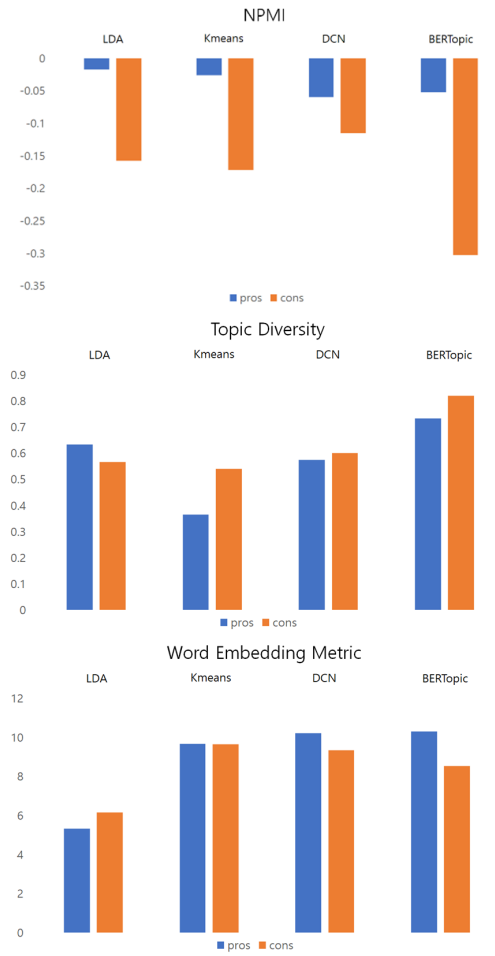


FIGURE 4. Topic modeling evaluation graphs.

Table 2 shows the specific example of feature extraction. The example product is a mold removal detergent. The analysis Identifies certain keywords associated with the product, namely 'mold,' 'cleaning,' 'odor,' and 'removal.' Based on the data presented in the table, it can be deduced that the product is effective in removing mold, but it also exhibits a significant issue with odor, which appears to be a prominent drawback. When devising a sales strategy for this product or a similar one, it would be prudent to focus efforts on mitigating the odor concern or highlighting its exceptional mold removal capabilities in advertising campaigns.

B. EXTRACTING RELATED PRODUCTS

To extract related products, we implemented topic modeling algorithms that rely on clustering. We evaluated the outcomes derived from three distinct models: Kmeans, DCN, and BERTopic. For this specific experiment, we standardized the number of topics to five.

DCN was trained using the same parameters as in the previous experiment, albeit with a reduced number of epochs (25) to circumvent issues related to trivial solutions. The extracted results are detailed in Table 3. We anticipate that training of models on a larger set of products and refined dataset, combined with a comprehensive interpretation of the generated topics, will provide more valuable insights.

The results show that interrelated products such as (male underwear, jeans, detergents) and (raincoats, crocs) are classified in the same cluster. Through this, it will be possible to help sell products by extracting products purchased with similar intentions and using them for product recommendations.

C. FORECASTING PRODUCT TRENDS

In this study, we aim to measure the performance of the model and compare it with other models. Initially, to investigate the influence of inputting product features into the model's decoder component on prediction, we compare the performance of the model without inputting product feature information into the decoder with that of the model that incorporates this information. Additionally, we evaluate the impact of a newly introduced loss function in this study by comparing it with a model using the popular MSE as the loss function. The entire model was trained for 200 epochs. Performance

TABLE 3. Results of extracting related products.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Kmeans	Men's underwear, Jeans, Detergent, Umbrella, Anti-reflux cushion	Clothes dryer, Slip-on shoes, Treadmill, Travel suitcase, Anti-reflux cushion	Interior decoration items, Key case, Crocs, Men's underwear, Jeans	Chicken breast, Crocs, Slip-on shoes, Anti-reflux cushion, Travel suitcase	Detergent, Jeans, Key case, Paraffin therapy device, Anti-reflux cushion
DCN	Men's underwear, Jeans, Detergent, Umbrella, Anti-reflux cushion	Chicken breast, Key case, Anti-reflux cushion, Travel suitcase, Paraffin therapy device	Detergent, Jeans, Key case, Paraffin therapy device, Men's underwear	Jeans, Interior decoration items, Chicken breast, Detergent, Clothes hanger	Clothes dryer, Slip-on shoes, Treadmill, Interior decoration items, Yoga mat
UMAP+HDBSCAN (BERTopic)	Interior decoration items, Chicken breast, Men's underwear, Crocs, Jeans	Detergent, Jeans, Raincoat, Key case, Clothes hanger	Men's underwear, Clothes hanger, Jeans, Travel suitcase, Paraffin therapy device	Men's underwear, Paraffin therapy device, Clothes hanger, Travel suitcase, Jeans	Raincoat, Men's underwear, Jeans, Umbrella, Crocs

evaluation metrics include MSE, Weighted Absolute Percentage Error (WAPE), and Pearson correlation coefficient. The performance measurement results are presented in Table 4.

TABLE 4. Results of trend forecasting.

	MSE	WAPE	Correlation
Product data: None Loss: MSE	12.992	48.412	0.291
Product data: topic modeling Loss: MSE	12.936	47.893	0.290
Product data: topic modeling Loss: MSE+Correlation	12.935	47.238	0.312
Product data: summarization Loss: MSE+Correlation	12.907	47.008	0.340

When observing the differences in performance based on whether product information is included, it is noted that the model that incorporates product information exhibited a marginal advantage. However, the correlation coefficient, which is used to measure the overall trend shape, did not show a significant difference.

Comparing the results of the two models with product information input to assess the impact of the loss function change, it is evident that the model incorporating the correlation with the loss function demonstrated superior performance, as expected. Moreover, it also maintained or even surpassed the performance of conventional error-based metrics such as MSE and WAPE. This observation indicates that the introduction of the correlation coefficient can help to reduce errors between data points, but also can contribute to capturing the overall trend of time series data. When comparing the methods of extracting product information, topic modeling and summarization, it was observed that summarization yielded slightly better results in terms of error-based evaluation metrics such as MSE and WAPE. In the case of correlation coefficients, a more substantial increase was evident when adding correlation coefficients to the loss function. This suggests that summarization, as compared to topic modeling, extracts data that is more closely related to the trends in the data.

To assess how helpful this model can be in trend analysis, we examined the distribution of calculated correlations

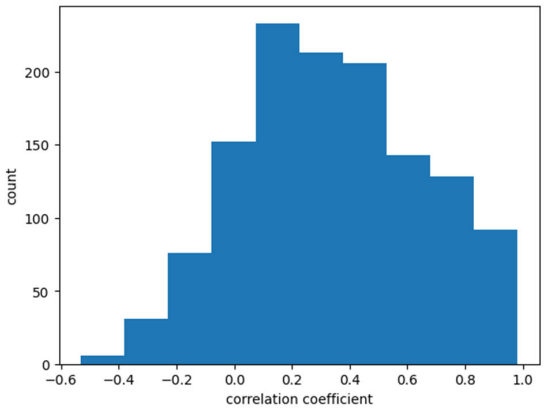


FIGURE 5. Distribution of correlation coefficients.

in the test dataset. The histogram representing this can be seen in Fig. 5. Overall, it is evident that a significant portion of the distribution is concentrated in the higher values. The proportion of data with positive correlation coefficients was approximately 86%, with data exceeding 0.3 at 53%, and data exceeding 0.5 at 32%. Given that it derives positive correlations in over 80% of the data, it appears that this model can provide sufficient assistance in identifying product trends. Additionally, data with negative correlation coefficients, when examined graphically, appears to sufficiently represent fluctuations over time or contains sudden increases that are challenging to predict based on historical data, as shown in Fig. 6. Hence, it is speculated that the model can be even more useful for its intended purpose than the actual numerical values suggest.

VI. DISCUSSION

In this study, we leveraged topic modeling to extract product characteristics and associated products from consumer reviews, with the aim of uncovering valuable marketing insights. During this process, the conventional clustering module of BERTopic was replaced with a recently developed neural network-based algorithm known as DCN. The primary objectives were to evaluate the outcomes of this task and to perform a comparative assessment of the performance of the DCN-based model against the baseline BERTopic and alternative clustering algorithms.

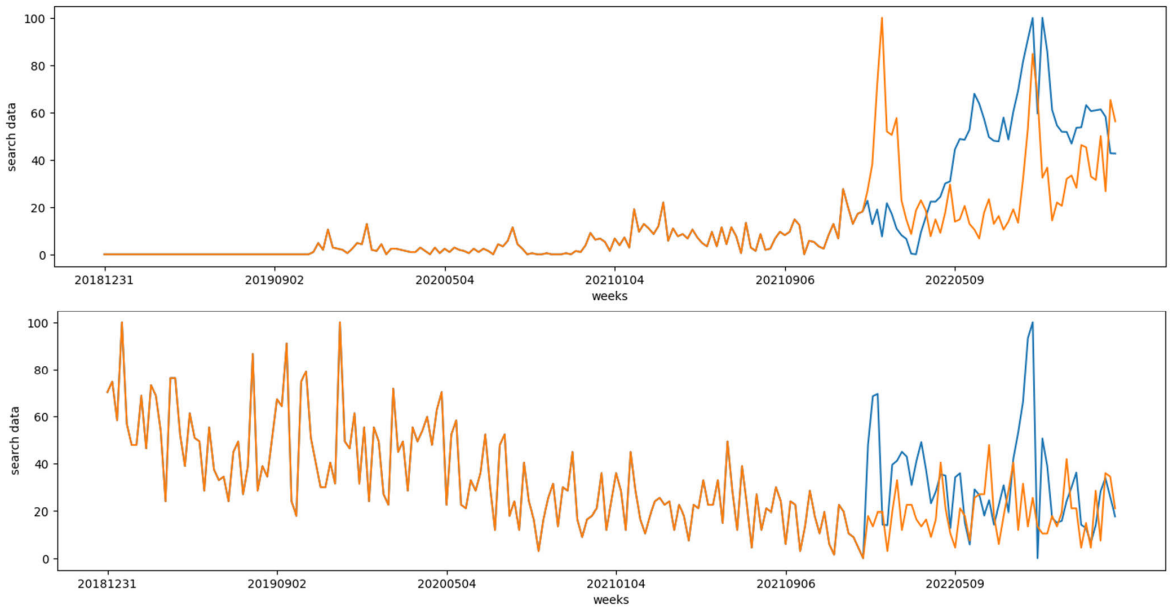


FIGURE 6. Example graph of low correlation coefficient data. (Yellow line: original data, blue line: forecasted data by model.)

The model implementing the DCN performed well, although its results were similar or slightly less favorable than the baseline BERTopic. This discrepancy can be attributed to fundamental disparities between the DCN and HDBSCAN. The DCN relies on the K-means algorithm, which centers on the distances between data points, while HDBSCAN employs data density for hierarchical clustering. The baseline BERTopic excludes reviews identified as noise during the clustering process, a practice anticipated to enhance its overall performance. However, it is important to note that reviews categorized as noise could contain valuable information, and misclassifying informative reviews as noise may result in the loss of significant insights. Thus, further investigation is warranted to address this aspect.

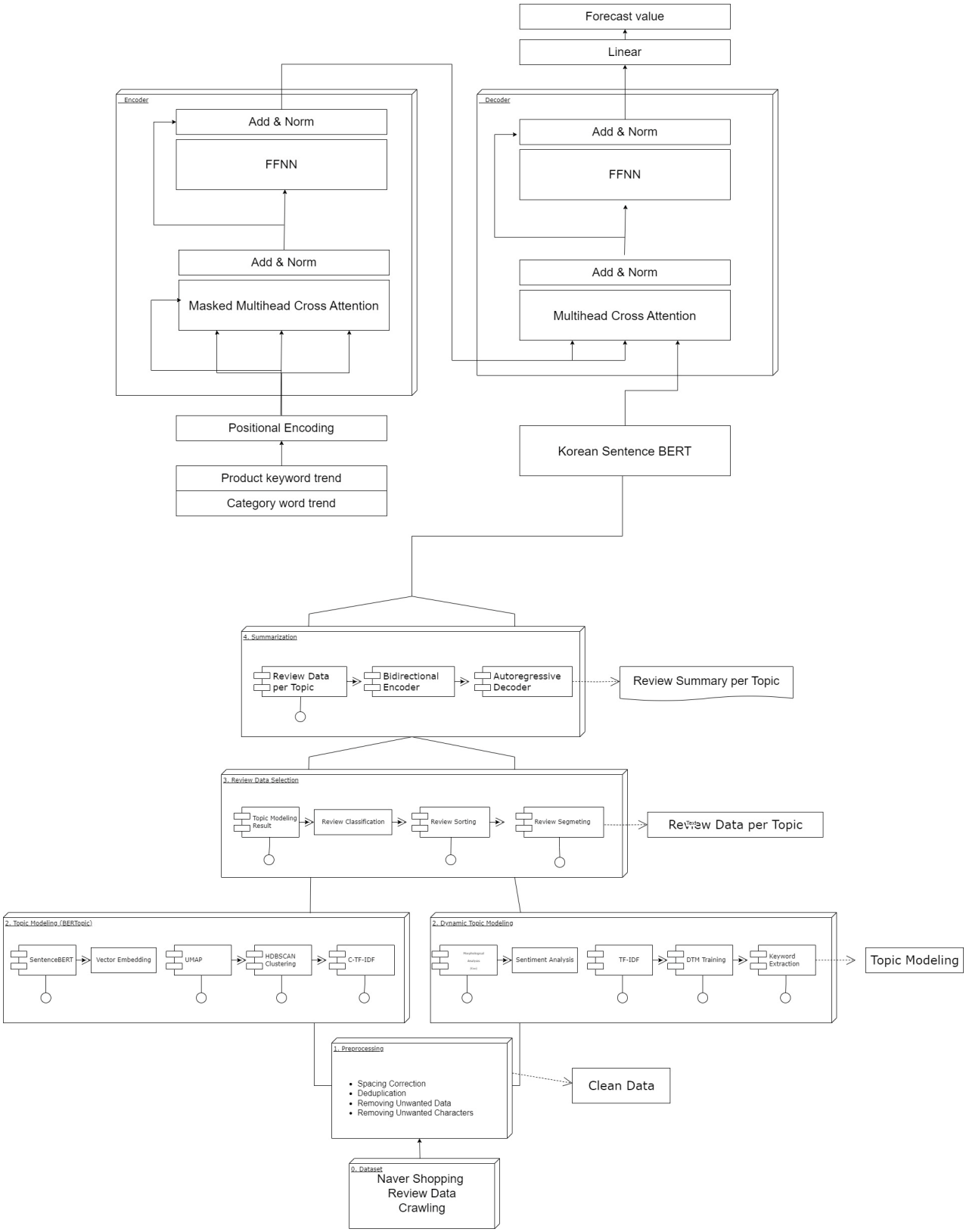
The product characteristics and related products extracted through the method proposed in this study have the potential to contribute to the discovery of marketing insights and the development of strategic approaches.

Insights derived from the examination of product characteristics can offer valuable information regarding consumer reactions and purchase intentions, thereby guiding strategies for product enhancement, emphasis in advertising, and more. These insights can serve as crucial reference materials in the formulation of marketing strategies. Regarding the extraction of related products, clustering items with similar content mentioned in reviews enables the identification of products that consumers purchase with similar intentions or features, even when they do not belong to the same product category. This information can inform decisions regarding the introduction of complementary products alongside high-selling items or the promotion of related products to consumers interested in specific products.

The product trend prediction model enhances its predictive performance by extracting key information about products from text data such as product reviews and product names through the application of topic modeling and TF-IDF. This information is then injected into the decoder of a Transformer model. Additionally, in the context of trend prediction, it is deemed more crucial to capture the overall upward or downward trend rather than minimizing numerical errors between two values. To address this, the correlation coefficient is incorporated into the loss function, thereby strengthening the correlation between predicted and actual values. This approach results in improved performance, not only in terms of error-based metrics such as Mean Squared Error (MSE) but also in capturing correlations. The trend prediction model proposed in this study can be of significant assistance in forecasting demand for specific products. This, in turn, enables sellers to maximize product sales by preparing products at specific times and proactively mitigate potential losses due to low sales volumes. However, this model has a limitation in that it is impossible to predict changes caused by other unexpected variables in the future because it predicts future data by identifying patterns of past data.

VII. CONCLUSION

The conducted topic modeling analysis provided insights into the characteristics and related products of the item. The evaluation of different models revealed that both DCN and BERTopic outperformed traditional algorithms in terms of overall performance. Topic modeling has also proven beneficial for trend prediction using time series data. When utilizing topic modeling to inject additional information about products, the predictive performance improved compared to using the conventional Transformer structure. Overall, the topic modeling approach proved to be effective in capturing the characteristics and related products of the chosen item, providing a deeper understanding of consumer preferences and market trends. This also confirms that topic modeling can be used directly or indirectly in addressing various problems.



APPENDIX

The figure above illustrates the overall structure of our framework, showcasing all the individual tasks within a single diagram. It involves collecting review data from an e-commerce platform through web scraping, providing the results of extracting product information using both topic modeling and summarization models. These results are further utilized to operate a transformer-based time series prediction model using search volume data, enabling the prediction of product trends.

REFERENCES

[1] Y. An, H. Oh, and J. Lee, "Marketing insights from reviews using topic modeling with BERTopic and deep clustering network," *Appl. Sci.*, vol. 13, no. 16, p. 9443, Aug. 2023.

[2] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/tit.1982.1056489.

[3] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science), vol. 7819, J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu, Eds. Berlin, Germany: Springer, 2013, doi: 10.1007/978-3-642-37456-2_14.

[4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, vol. 96, 1996, pp. 226–231.

[5] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 478–487.

[6] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 3861–3870.

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

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

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[10] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[11] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11106–11115.

[12] H. Wu, J. Xu, J. Wang, and M. Long, "AutoFormer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22419–22430.

[13] M. Caron and O. Müller, "Hardening soft information: A transformer-based approach to forecasting stock return volatility," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 4383–4391, doi: 10.1109/BigData50022.2020.9378134.

[14] G. Skenderi, C. Joppi, M. Denitto, and M. Cristani, "Well Googled is half done: Multimodal forecasting of new fashion product sales with image-based Google Trends," 2021, *arXiv:2109.09824*.

[15] V. S. Anoop and S. Asharaf, "A topic modeling guided approach for semantic knowledge discovery in e-commerce," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 4, no. 6, p. 40, 2017.

[16] W. Li, J. Yin, and H. Chen, "Supervised topic modeling using hierarchical Dirichlet process-based inverse regression: Experiments on e-commerce applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1192–1205, Jun. 2018, doi: 10.1109/TKDE.2017.2786727.

[17] R. Chen and W. Xu, "The determinants of online customer ratings: A combined domain ontology and topic text analytics approach," *Electron. Commerce Res.*, vol. 17, no. 1, pp. 31–50, Mar. 2017, doi: 10.1007/s10660-016-9243-6.

[18] L.-Y. Dong, S.-J. Ji, C.-J. Zhang, Q. Zhang, D. W. Chiu, L.-Q. Qiu, and D. Li, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Syst. Appl.*, vol. 114, pp. 210–223, Dec. 2018.

[19] P. Hajek, L. Hikkerova, and J.-M. Sahut, "Fake review detection in e-commerce platforms using aspect-based sentiment analysis," *J. Bus. Res.*, vol. 167, Nov. 2023, Art. no. 114143.

[20] X. Ye, Z. Lian, B. She, and S. Kudva, "Spatial and big data analytics of e-market transactions in China," *GeoJournal*, vol. 85, pp. 329–341, Jan. 2020.

[21] W. Hong, C. Zheng, L. Wu, and X. Pu, "Analyzing the relationship between consumer satisfaction and fresh e-commerce logistics service using text mining techniques," *Sustainability*, vol. 11, no. 13, p. 3570, Jun. 2019.

[22] Y. D. Wang, B. T. Jiang, and X. Y. Ye, "A method for studying the development pattern of urban commercial service facilities based on customer reviews from social media," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 41, pp. 577–578, Jun. 2016.

[23] M.-G. Jeong, J.-Y. Kwon, J.-W. Lee, Y.-N. Lee, and S.-B. Lee, "Text mining analysis of consumer perception of food distribution platforms : Focusing on topic modeling," *Foodservice Manage. Soc. Korea*, vol. 24, no. 7, pp. 71–100, Nov. 2021.

[24] B. Lee and H. Noh, "Food tourism market segmentation approach using topic modeling analysis: Focusing on benefits sought," *Korean J. Hospitality Tourism*, vol. 29, no. 4, pp. 187–204, Jun. 2020.

[25] L. Soyeon and K. Yeongok, "Analysis of apartment interior trend using topic modeling: Focusing on 'today's house' review data," in *Proc. 14th Int. Conf. Knowl. Manag. Inf. Syst.*, Valletta, Malta, Oct. 2022, pp. 141–149.

[26] M.-K. Cho and B.-J. Lee, "Comparison of service quality of full service carriers in Korea using topic modeling: Based on reviews from TripAdvisor," *J. Hospitality Tourism Stud.*, vol. 23, no. 1, pp. 152–165, Feb. 2021.

[27] V. Ekambaram, K. Manglik, S. Mukherjee, S. S. K. Sajja, S. Dwivedi, and V. Raykar, "Attention based multi-modal new product sales time-series forecasting," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3110–3118.

[28] T. Kudo. (2006). *Mecab: Yet Another Part-of-Speech and Morphological Analyzer*. [Online]. Available: https://sourceforge.net/projects/mecab/

[29] L. Junbum, "KcBERT: Korean comments BERT," in *Proc. 32nd Annu. Conf. Hum. Cogn. Lang. Technol.*, 2020, pp. 437-440.

[30] T. R. Derrick, B. T. Bates, and J. S. Dufek, "Evaluation of time-series data sets using the Pearson product-moment correlation coefficient," *Med. Sci. Sports Exerc.*, vol. 26, no. 7, pp. 919–928, Jul. 1994.

[31] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proc. GSCL*, vol. 30, 2009, pp. 31–40.

[32] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 439–453, Dec. 2020.

[33] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Using word embedding to evaluate the coherence of topics from Twitter data," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 1057–1060.

[34] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, and A. Candelieri, "OCTIS: Comparing and optimizing topic models is simple!" in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Syst. Demonstration*, 2021.

YUSUNG AN is currently pursuing the degree with the College of Computing and Informatics, Sungkyunkwan University, South Korea. His research interests include data analysis and natural language processing techniques using machine learning and deep learning.

DONGJU KIM was born in Ansan-si, Gyeonggi-do, South Korea, in 2003. She is currently enrolled with the Department of Software Engineering, Sungkyunkwan University. In Summer 2023, she participated in a collaborative industry-academia project under the auspices of the Software Engineering College, Sungkyunkwan University.

JUYEON LEE was born in Seoul, in March 2003. Since April 2023, she has been actively involved in the development of a marketing engine based on big data analysis and natural language processing with MechaSolution. She is currently a Sophomore with the Department of Computer Science and Engineering, Sungkyunkwan University, South Korea.

HAYOUNG OH received the Ph.D. degree in computer science from Seoul National University, Seoul, South Korea, in 2013.

From 2001 to 2004, she was involved in research and development with the Institute of Shinhan Financial Group, Seoul. She was a Visiting Scholar with U.C. Berkeley, in 2010. She was an Assistant Professor of computer science with Soongsil University and Ajou University, from 2014 to 2019. Since 2020, she has been an Associate Professor with the College of Computing and Informatics, Sungkyunkwan University. Her major was artificial intelligence with big data analysis. Her research interests include social network analysis, recommender systems, spam detection, and natural language processing techniques using deep learning and big data analysis.

JOO-SIK LEE is currently a Professor with the College of Computing and Informatics, Sungkyunkwan University, South Korea. He has conducted many studies on multi-modal communications, platform services, broadcasting, and the applications of artificial intelligence and deep learning.

DONGHWA JEONG is currently the CEO of MechaSolution, South Korea. He is also the Leader of securing price competitiveness through OEM. MechaSolution strives for customer satisfaction with low prices through custom-made manufacturing methods. In addition, it is taking the lead in maximizing the synergistic effects of industry-university cooperation. Activation of small- and medium-sized enterprises (SMEs) is based on the advancement of professional (specialized) education through industry-university cooperation from the course and improvement of the technology. The company is working to revive the local economy by realizing and commercializing innovative and creative ideas using machine learning and deep learning.