applied sciences

MDPI

Article

Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network

Yusung An, Hayoung Oh * and Joosik Lee

College of Computing and Informatics, Sungkyunkwan University, Seoul 16419, Republic of Korea; dksdbtjd2@gmail.com (Y.A.); jsleeas@gmail.com (J.L.)
* Correspondence: hyoh79@gmail.com

Abstract: The feedback shared by consumers on e-commerce platforms holds immense value in marketing, as it offers insights into their opinions and preferences, which are readily accessible. However, analyzing a large volume of reviews manually is impractical. Therefore, automating the extraction of essential insights from these data can provide more comprehensive and efficient information. This research focuses on leveraging clustering algorithms to automate the extraction of consumer intentions, related products, and the pros and cons of products from review data. To achieve this, a review dataset was created by performing web crawling on the Naver Shopping platform. The findings are expected to contribute to a more precise understanding of consumer sentiments, enabling marketers to make informed decisions across a wide range of products and services.

Keywords: clustering; marketing; topic modeling

Citation: An, Y.; Oh, H.; Lee, J. Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network. *Appl. Sci.* **2023**, *13*, 9443. https://doi.org/10.3390/app13169443

Academic Editors: Wenfeng Zheng, Mingzhe Liu, Kenan Li and Xuan Liu

Received: 9 July 2023
Revised: 17 August 2023
Accepted: 17 August 2023
Published: 21 August 2023

1. Introduction

In the digital age, consumers have increasingly been leaving and sharing reviews about various products and services through online platforms. These review data have become a valuable asset for companies as the data can be used to understand consumer opinions and preferences, thereby improving marketing strategies. However, effectively analyzing a large volume of review data and extracting useful information from it is a challenging task.

This research proposes a method of extracting and analyzing marketing-related information based on review data using clustering-based topic modeling as the core technique. A clustering algorithm is a computational method used in data analysis and machine learning to group a set of data points into clusters based on their similarities. Topic modeling is a natural language processing technique used to extract hidden topics from text data. Text data can contain various topics, and each document may be related to multiple topics. Topic modeling automatically identifies these topics and estimates which topics each document is composed of.

Topic modeling and clustering algorithms are a methodology widely used in marketing, such as for customer segmentation and recommendation systems, as they group data with similar characteristics. In a study [1], product description data were collected from an e-commerce platform that sells mobile phones, and topic modeling was conducted using LDA, a traditional topic modeling method. 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 find semantic knowledge about the products. LDA has also 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, to analyze consumer opinions [2–5].

In this study, we introduce a process that utilizes clustering algorithms based on review data that are accessible from external sources to analyze consumer intent and extract the relationship between products. By clustering review data left by consumers, groups

with similar consumer opinions can be formed, enabling the identification of consumer purchase intent and the relevance between products. This provides important information for companies to understand the association between specific products and other products, allowing them to develop sales strategies. Additionally, we present a process to extract the advantages and disadvantages of products from review data. Using topic modeling, we classify review data into respective topics and extract the advantages and disadvantages of products from each topic. This allows companies to identify which products consumers evaluate most positively and areas that need improvement. It can serve as valuable reference material for product development and marketing strategy formulation.

There are a variety of topic modeling algorithms, but this study focuses on BERTopic [6]. BERTopic demonstrates a better performance as it leverages BERT [7] for vector embedding and considers the semantics and contextual information of the text. BERTopic is a topic modeling approach that utilizes BERT's embedding ability and demonstrates that the task can be addressed simply by employing a clustering-based approach. In this study, other clustering algorithms with good performance, such as the DCN, have been used as alternatives to generate results and compare their performances. By incorporating different clustering algorithms into the basic structure of BERTopic, this research aims to explore and compare the results obtained. This will provide insights into the effectiveness and performance of various clustering algorithms in the context of topic modeling.

2. Background Algorithms

2.1. K-Means Algorithms

One of the most representative clustering algorithms is the K-means algorithm [8]. The K-means algorithm selects k cluster centroids and iteratively moves them to optimize the following cost function during training:

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

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.

2.2. HDBSCAN

There is another algorithm called hierarchical clustering. It is a bottom-up approach that creates a hierarchy of clusters by iteratively merging or splitting data points based on their similarities. The basic idea behind hierarchical clustering is to start with each data point as its own cluster and then iteratively merge or split clusters until a stopping criterion is met. The result is a hierarchical structure that is often represented as a dendrogram, which shows the relationships between clusters at different levels of similarity. In this study, we will use the HDBSCAN algorithm [9] among various hierarchical clustering algorithms. This algorithm is a modification of the DBSCAN algorithm [10] that is a part of the hierarchical algorithm paradigm. It calculates the mutual reachability between each pair of data points and constructs a minimum spanning tree based on these points to transform it into a hierarchical structure for clustering. This approach has the advantage of being able to find clusters of various shapes better than the K-means algorithm. One important factor to consider when performing clustering is the dimensionality of the data. If the dimensionality of the data is too high, it becomes difficult to achieve good clustering results due to the curse of dimensionality. Therefore, in cases where the dimensionality of the data is excessively high, it is common to incorporate dimensionality reduction algorithms such as UMAP into the clustering process.

2.3. Deep Clustering Network (DCN)

One of the other methods for reducing the dimensionality of the data is to utilize autoencoders to create latent vectors of the data. The algorithm initially proposed for this

purpose is called Deep Embedded Clustering (DEC) [11]. This algorithm utilizes stacked autoencoders to extract lower-dimensional latent vectors for the input data, which are then used for clustering, demonstrating good performance. The Deep Clustering Network (DCN) improves upon DEC by modifying the loss function to achieve slightly better performance. This method also relies on autoencoders for dimensionality reduction and utilizes them for clustering [12]. The novelty of this research lies in the definition of a new loss function by combining the loss functions of autoencoders and the K-means algorithm. This new loss function is used during the dimensionality reduction process to generate vectors that are more suitable for the K-means algorithm, resulting in enhanced performance for clustering. Below is the expression for the loss function used in this research.

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

The term on the left within the sigma represents the loss function of the autoencoder, which calculates the error between the input data and the reconstructed output data. In this research, the Mean Square Error (MSE) was used, but other functions, such as L1-norm or KL-divergence, can also be used. On the right is the form of the loss function for the K-means algorithm. By attaching a coefficient lambda to it, the degree of adjustment and influence can be controlled.

2.4. Latent Dirichlet Allocation (LDA)

The most widely used technique among various topic modeling methods is Latent Dirichlet Allocation (LDA) [13]. LDA assumes that documents are composed of multiple topics, each document is a mixture of topics, and topics are represented by word distributions. It iteratively infers the topic distribution within documents and the word distribution within topics. While LDA performs well and is widely used, it has limitations in capturing semantic relationships between words.

2.5. BERTopic

One topic modeling method that overcomes these limitations is BERTopic, which, as the name suggests, utilizes BERT [7]. It first embeds the text data into vectors using Sentence BERT [14] and then performs clustering on these vectors to conduct topic modeling. Clustering is carried out by reducing the dimensionality of the vectors using UMAP, as mentioned before, and by applying HDBSCAN. BERTopic demonstrates a better performance as it leverages BERT for vector embedding and considers the semantics and contextual information of the text. BERTopic is a topic modeling approach that utilizes BERT's embedding ability and demonstrates that the task can be addressed simply by employing a clustering-based approach. In this study, in addition to HDSCAN, which is the default clustering algorithm in BERTopic, other clustering algorithms with a good performance, such as the DCN, have been used as alternatives to generate results and compare their performances. By incorporating different clustering algorithms into the basic structure of BERTopic, this research aims to explore and compare the results obtained. This will provide insights into the effectiveness and performance of various clustering algorithms in the context of topic modeling.

3. Literature Review

Efforts to uncover latent insights through the utilization of topic modeling have been present across various fields, including marketing. In a previous study [15], the factors that impacted the user experience on mobile health apps were analyzed using BERTopic with registered user review data, providing insights into the direction of service development. Additionally, basic topic modeling methods have been expanded through the incorporation of new models or fusion with other machine learning models. Study [16] introduced a supervised topic modeling approach, Hierarchical Dirichlet Process-based Inverse Regres-

sion (HDP-IR), which is based on LDA and Inverse Regression, to extend to the approach's use in predictive models for variables such as customer sentiment, product quality, and affect. In study [17], BERTopic was used to extract topic probability distributions from topic modeling results from a Twitter dataset. These distributions were then combined with vector representations of the original data for use in a classification model, which demonstrated a better performance than a basic classification model. Study [18] proposed a novel robust risk maximum expert consensus model by combining the mean-variance theory with the traditional maximum expert consensus model. This model enables the derivation of consensus decisions for the entire group from individual opinions, and it can also be employed to analyze consumer opinions embedded within review data.

Furthermore, there have been variations in the configuration of BERTopic. Study [19] compared the performance of BERTopic's clustering module by switching from HDBSCAN to the K-means algorithm. The results measured using the representative evaluation metrics of topic modeling, such as NPMI and Topic Diversity, showed that the model incorporating the K-means algorithm did not surpass the performance of the baseline model in several benchmarks and datasets.

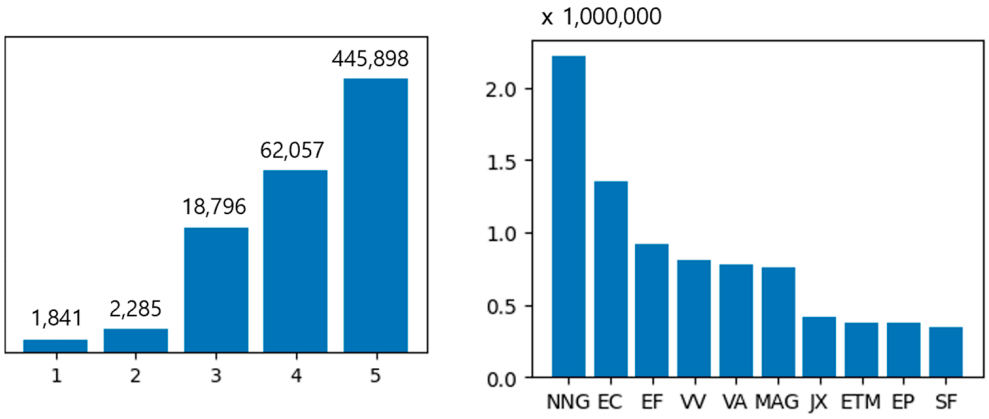
4. Methodology

In this section, based on the consumer review data collected from an actual e-commerce platform, we will explain the method of (1) the process of extracting the advantages and disadvantages of each product and (2) the process of finding the relationships between the products. These processes will utilize clustering algorithms and topic modeling techniques such as BERTopic and the DCN, and we will also compare their performance with traditional methods such as LDA and K-means algorithms. The data crawling, implementation, and experimentation of all models were carried out using Python. The basic BERTopic was utilized using the provided library, while the DCN-based model was adapted from the official implementation code by the DCN researchers to suit the data of this study. The K-means algorithm employed the code from the sklearn library [20], and LDA as well as the evaluation metric calculations, were performed using OCTIS [21].

4.1. Data Collection and Preprocessing

We chose 'Naver Shopping', one of the largest e-commerce platforms in Korea, as the platform to collect consumer reviews. We selected 17 product keywords based on sales scale and crawled the review content and ratings for each product, resulting in a total of approximately 500,000 reviews. The collected reviews were saved and managed in a csv file format. Figure 1a illustrates the distribution of star ratings for the crawled reviews. It is evident that the majority of reviews are assigned 5 stars, and as the star ratings decrease, the number of reviews also decreases. Due to significant variations in the number of reviews between star ratings, a logarithmic scale was employed on the y-axis for visualization purposes.

The collected data were preprocessed using the Korean corpus-trained morphological analyzer Mecab-ko, which decomposed the data into morphemes. Based on our intuition and observations, we determined that important information about the products is concentrated in nouns, verbs, and adjectives in the review text. Therefore, we kept only these three parts of speech as morphemes and removed all others. We also removed additional Korean stopwords. Figure 1b presents the frequencies of the top 10 most frequently occurring parts of speech (POS) after morphological analysis of all reviews. Generally, nouns, verbs, adjectives, and adverbs are observed to occur frequently. This observation serves as another basis for extracting and utilizing only these four POS categories in the preprocessing of the study.



(**a**) (**b**)

Figure 1. (**a**) Distribution of collected review data star ratings (1~5). (**b**) 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 particle, SF: sentence-finishing punctuation marks.).

4.2. Extracting Advantages and Disadvantages

To extract the advantages and disadvantages, we divided the review data based on the star ratings. We used only the reviews that received a 5-star rating for extracting advantages and reviews that received a 1-star or 2-star rating for extracting disadvantages. We then proceeded with topic modeling using these segmented review data.

Next, we embedded the review data into vectors. In this process, we utilized KcBERT [22], which is a BERT model trained using comments collected from Korean online news. KcBERT performs well in natural language processing tasks for the informal text that often contains colloquial language, slang, and typographical errors. We employed mean pooling with KcBERT to embed the review data into vectors.

After embedding the data, we conducted topic modeling. We hypothesized that the keywords extracted from the topics would effectively represent the characteristics (advantages and disadvantages) of the products. We then explored the BERTopic method and a new approach using the DCN module as a replacement for BERTopic's clustering module. The overall process is depicted in Figure 2.

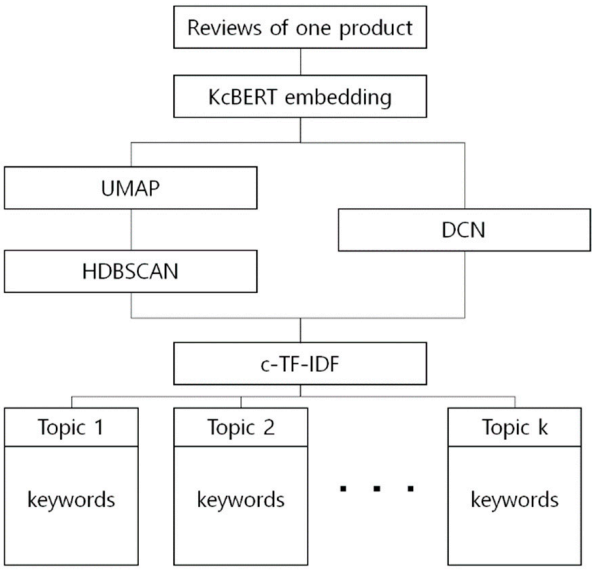


Figure 2. Process of extracting features of a product from a review.

4.3. Analysis of Related Products

The analysis of related products follows a process similar to the one described for extracting advantages and disadvantages. The main difference is that while extracting advantages and disadvantages, we performed clustering-based topic modeling for individual product reviews. In the analysis of related products, we conducted topic modeling using the reviews of multiple products. We then examined which product reviews appear together within each topic. If two products are frequently co-located within the same topic, this suggests that they share certain characteristics and indicates a relationship between the two products. Therefore, when extracting product features, we focused on the frequency of co-occurrence of the product by using TF alone instead of using c-TF-IDF to extract products from clustered topics. The overall process is depicted in Figure 3.

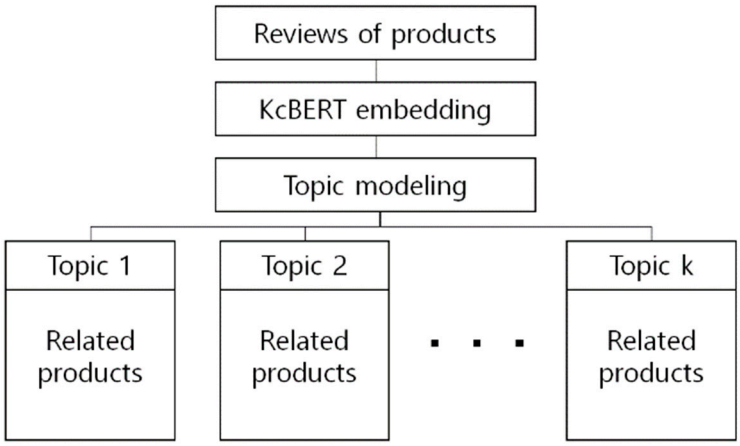


Figure 3. Process of extracting related products.

4.4. Hyperparameter Optimization

Both clustering algorithms and topic modeling are highly influenced by an important initial hyperparameter: the number of clusters or topics. One widely known method, especially used in K-means-based clustering algorithms, is the elbow method [23]. This method involves calculating the Within Clusters Sum of Squares (WCSS) metric for different numbers of clusters and selecting the point on the WCSS cluster number graph where the curve resembles an elbow shape as the optimal number of clusters. Another widely used method is the silhouette score, which measures how similar an object is to its own cluster compared to other clusters [24].

However, in this study, as clustering is used as part of the topic modeling process, we aimed to optimize the hyperparameters based on the results of the topic modeling. There are various metrics available to evaluate the results of topic modeling [25]. In this study, we utilized the coherence of topics metric based on word embeddings [26]. The formula for this metric is as follows:

$$W_{CE} = \frac{1}{k}\sum_{t=1}^{k}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} cosine\_similarity(w_{ti}, w_{tj})$$

where k is the number of topics, N is the number of keywords extracted from topic clusters, and w_{ti}, w_{tj} is each keyword's embedding vector within topic t. Our goal is to extract product characteristics from the keywords extracted from each topic. Thus, the more consistent the extracted keywords are within each topic, the more they align with our objective. Therefore, the hyperparameter optimization process was conducted by changing the number of topics (clusters) and selecting the number that resulted in the highest value of the word embedding metric. We used the Korean pretrained FastText model [27]. In the authors' original paper, GloVe was used, but due to the nature of the Korean language, errors in the morphological analysis are not negligible, and Out of Vocabulary (OOV) issues can

occur frequently. FastText, on the other hand, considers the subwords of words and generates embedding vectors, making it capable of addressing OOV problems, unlike other word embedding methods. Therefore, considering the characteristics of our current task and the potential occurrence of OOV problems, we chose FastText as it is more suitable.

5. Experiments

In this section, we will examine the process of extracting product advantages, disadvantages, and related products using real data. Then, we will compare the performance of BERTopic and the 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 topics determined through the hyperparameter optimization process mentioned earlier. For the DCN, the training process involved two stages: pretraining using stacked autoencoders and actual DCN training. Both stages were trained for 50 epochs. However, we encountered a problem with the trivial solution due to a setting in the original research when using the MNIST benchmark dataset, where all data were 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 because the original value appeared excessively large in comparison to the embedding vector dimension of 768 employed in this experiment. This discrepancy raised concerns akin to the curse of dimensionality, suggesting that the characteristics of each vector were being overlooked, resulting in potential issues. The dimension of the generated latent vectors was set to 10.

In the task of extracting the pros and cons of products, we conducted experiments using traditional topic modeling methods such as LDA, the default configuration of BERTopic, and modified models of BERTopic with its clustering module replaced by the K-means algorithm and DCN. Additionally, for the task of extracting related products, since the clustering process is essential, we utilized K-means, the DCN, and the modified BERTopic model but excluded LDA.

6. Results

6.1. Extracting Features

The performance of the models was evaluated using three metrics. Topic coherence is an evaluation metric derived from the assumption that better-performing topic models will result in extracted keywords for each topic that have semantically similar meanings. In this experiment, we used NPMI (Normalized Pointwise Mutual Information) [28] and Word Embedding-based Topic Coherence. NPMI quantifies how often two words appear together in a text. It ranges from -1 to 1, with higher values indicating better coherence. Word Embedding Coherence (WE), previously mentioned in Section 4.4, is an evaluation metric that involves embedding the extracted keywords as vectors and calculating the similarity between words. A higher value indicates a higher semantic similarity among keywords within the topic. Topic Diversity (TD) is an evaluation metric that assesses how distinct the extracted topics are from each other in terms of meaning [29]. It measures how different semantically meaningful keywords have been extracted among different topics. It is calculated as the proportion of unique words across all topics. It ranges from 0 to 1, with values closer to 1 indicating a higher diversity. The results of the topic modeling for a randomly selected product, which is detergent, are presented in Figure 4.

The left side of each graph represents the result evaluation metric for extracting advantages, while the right side represents the evaluation metric for extracting disadvantages.

The extraction of disadvantages received lower scores compared to the extraction of advantages in terms of performance. This is likely due to the difference in dataset sizes. Generally, the number of reviews with low ratings is significantly lower than those with

high ratings in consumer-generated reviews. Therefore, the amount of data used for extracting disadvantages was much smaller in the collected dataset, which is reflected in the results.

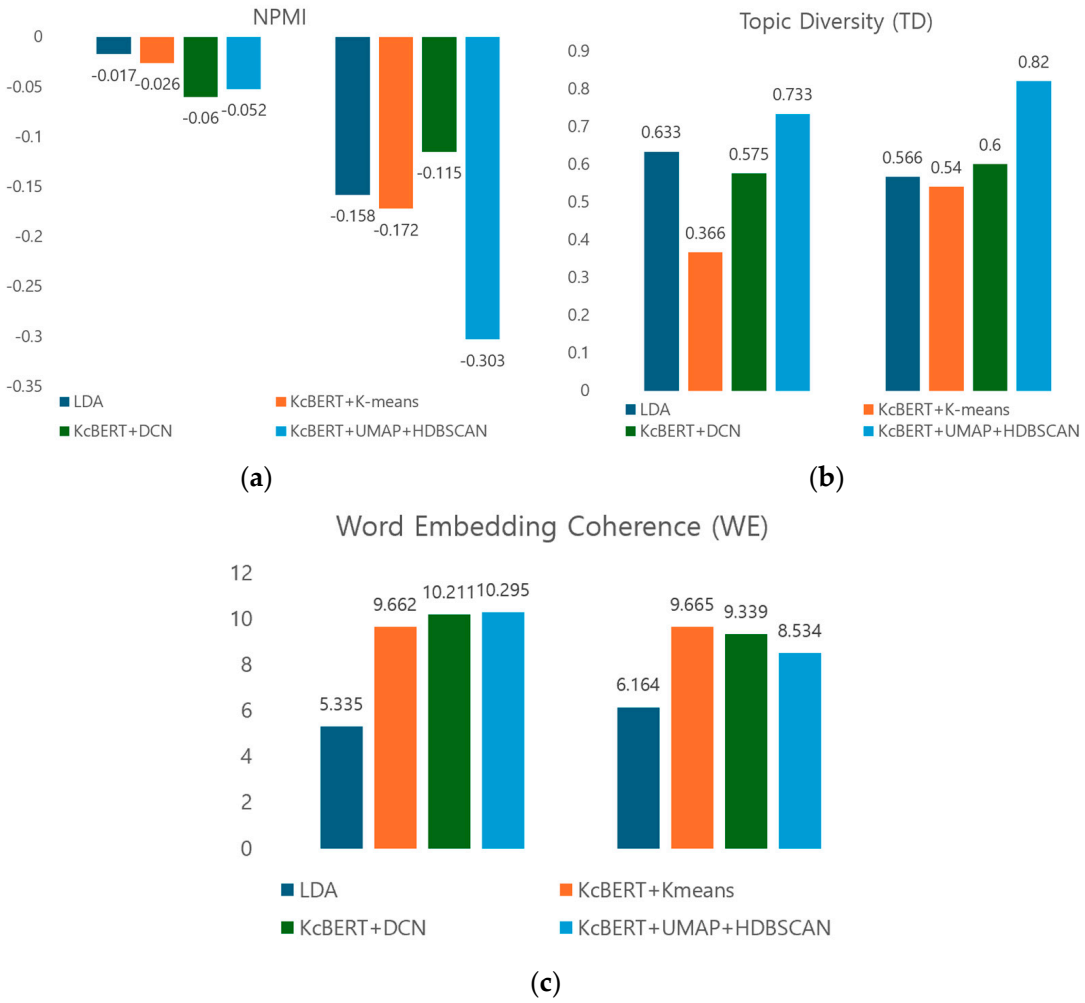


Figure 4. Topic modeling evaluation. (**a**): NPMI, (**b**): Topic Diversity, (**c**): WE Coherence. KcBET + UMAP + HDBSCAN equals the basic BERTopic.

When comparing the performance among the models, the more recent algorithms using the DCN and BERTopic generally outperformed traditional algorithms. Between the DCN and BERTopic, BERTopic showed slightly higher scores. Particularly, BERTopic exhibited the most outstanding performance in terms of Topic Diversity among the four models. However, for the word embedding metric, which was emphasized in this study, the DCN also demonstrated a good level of performance comparable to BERTopic. An example of extracted features of a product is shown in Table 1. The keywords in the table below were extracted from two or more of the various models, and the ones that have marketing significance were selected. The original topic modeling results can be found in Appendix A.

The selected product is a mold remover detergent. We can see that keywords related to the product include 'mold', 'cleaning', 'odor', and 'removal'. From the table, it can be inferred that the product has a good effect on mold removal but has an intense odor, which seems to be its major drawback. If devising a sales strategy for this product or a similar one, efforts could be directed towards mitigating the odor issue or emphasizing its effective mold removal capabilities by crafting advertisements.

Table 1. Results of extracting features of a product.

Pros		Cons	
Keyword	**Models**	**Keyword**	**Models**
mold	All	odor	All
removal	All	bleach	All
cleaning	All	intense	DCN BERTopic
toilet	All	same	K-means BERTopic
spray	All		
bathroom	K-means DCN BERTopic		

Furthermore, in the case of the LDA and K-means methods, which exhibited relatively low topic diversity values, the occurrence of words such as 'mold', 'spray', 'use', and 'removal' across multiple topics with overlaps can be observed. On the contrary, for the DCN-based model and BERTopic, which had relatively high topic diversity values, there were fewer overlapping words between different topics and a greater diversity of extracted keywords. This could potentially help reduce overlooked information in consumer reviews and yield more insights. In terms of topic coherence, higher values facilitated the comprehension of what content described the entire topic. For instance, in the case of LDA with the lowest WE value, words such as 'mold', 'delivery', 'product', and 'price', which are semantically distant, appeared together within one topic, making it challenging to grasp the overarching theme. Conversely, the DCN-based model and BERTopic exhibited favorable values, enabling easier inference of the predominant content within each topic. As an illustrative example from the advantages extracted by BERTopic, the first topic predicted the dominance of 'mold removal effectiveness', the second topic represented 'bathroom/cleaning', and the third topic pertained to 'wall cleaning/price discounts'. Therefore, we believe that topic coherence, when used alongside the size information of categorized topic clusters, could serve as a significant metric when analyzing prevalent opinions and more in-depth insights.

6.2. Extracting Related Products

In the process of extracting related products, we used clustering-based topic modeling algorithms. Therefore, we examined the results obtained using three models: K-means, the DCN, and BERTopic. In this experiment, we unified the number of topics to be 5. The DCN was trained with the same settings as the previous experiment but with a reduced number of epochs (25) to prevent trivial solution issues. The extracted results are presented in Table 2. It is anticipated that by training the models with a larger set of products and refined data and interpreting the generated topics together, more valuable information can be obtained.

Table 2. Results of extracting related products.

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
K-means	Men's underwear	Clothes dryer	Interior items	Chicken breast	Detergent
	Jeans	Slip-on shoes	Key case	Crocs	Jeans
	Detergent	Treadmill	Crocs	Slip-on shoes	Key case
	Umbrella	Travel suitcase	Men's underwear	Anti-reflux cushion	Paraffin machine
	Anti-reflux cushion	Anti-reflux cushion	Jeans	Travel suitcase	Anti-reflux cushion

Table 2. *Cont.*

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

7. Discussion

In this study, topic modeling was employed to extract product characteristics and related products from consumer reviews for potential utilization in marketing insights. During this process, the clustering module of BERTopic was replaced with a DCN, a recently created neural network-based algorithm, to examine the outcomes of the task. Furthermore, a performance comparison was conducted between the DCN-based model, the baseline BERTopic, and other alternative algorithms. Despite the various clustering algorithms that can be applied to BERTopic's clustering module, instances of utilizing recently proposed neural network-based clustering algorithms were limited. The endeavor to incorporate the DCN carries significance from the perspective of applying unexplored clustering algorithms to BERTopic and comparing their performances.

The model employing the DCN exhibited a favorable performance; however, in comparison to the baseline BERTopic, it showed similar or slightly inferior results. This disparity might be attributed to the fundamental paradigm differences between the DCN and HDBSCAN. The DCN relies on the K-means algorithm, which utilizes distances between data points, while HDBSCAN employs data density for hierarchical clustering. The baseline BERTopic benefits from excluding reviews identified as noise during the clustering process, which is anticipated to positively impact its performance. Nevertheless, reviews categorized as noise could potentially be covered to some extent during data preprocessing, and there is a possibility of the loss of valuable information if informative reviews are mistakenly categorized as noise. Therefore, further investigation is warranted in this regard.

The extracted product characteristics and related products obtained by utilizing the proposed method in this study can potentially aid in discovering marketing insights and devising strategies. Observing the extracted product characteristics can provide insights into consumer reactions and purchasing intentions, informing strategies for product improvement, advertising emphasis, and more. The insights gained can be utilized as reference materials for formulating marketing strategies. Concerning the extraction of related products, clustering products with similar content mentioned in reviews allows for the identification of products that consumers purchase with similar intentions or features, even if they do not belong to the same category. This information can aid in decisions about which additional products to offer alongside high-selling items or in exposing related products to consumers purchasing a specific item.

8. Conclusions

Demonstrating the potential utility of identifying topic-modeling-derived pros and cons of products and related products, this study underscores the capacity of using these data to facilitate the formulation of marketing strategies for product sales. Moreover, the

comparison of performance results by substituting BERTopic's clustering module with the DCN, a recently created neural network algorithm, exhibits the favorable performance of the DCN-based model and the baseline BERTopic. Furthermore, the exploration of various clustering algorithms suggests the possibility of enhancing the performance of baseline BERTopic across multiple tasks. However, it is noted that models targeting low-rated reviews, which commonly exhibit a limited data volume, exhibited a comparatively lower performance, necessitating further investigations to address this limitation.

Author Contributions: Conceptualization, Y.A., H.O. and J.L.; methodology, Y.A.; software, Y.A.; validation, Y.A., H.O. and J.L.; formal analysis, Y.A. and H.O.; investigation, Y.A.; resources, Y.A.; data curation, Y.A.; writing—original draft preparation, Y.A.; writing—review and editing, Y.A., H.O. and J.L.; visualization, Y.A.; supervision, H.O. and J.L.; project administration, J.L.; funding acquisition, H.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1F1A1074696), the Technology Innovation Program (or Industrial Strategic Technology Development Program-Source Technology Development and Commercialization of Digital Therapeutics) (20014967, Development of Digital Therapeutics for Depression from COVID-19) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), AI Convergence Research Fund, Sungkyunkwan University, 2023 and SMC-SKKU Future Convergence Research Program Grant.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to license and security issues.

Acknowledgments: The authors would like to express their gratitude to CEO Donghwa Jeong of Mechasolution for proposing a problem that needed to be solved and providing research opportunities.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The table below shows the extracted advantages and disadvantages of the product (mold remover detergent) for each model. Each row represents a keyword extracted for each topic. For example, in the case of LDA, both the advantage extraction and disadvantage extraction consist of three topics.

	Pros	Cons
LDA	곰팡이 (mold)	락스 (bleach)
	제거 (removal)	냄새 (odor)
	배송 (delivery)	곰팡이 (mold)
	빠르 (fast)	제거 (removal)
	지워 (erased)	사용 (use)
	만족 (satisfaction)	청소 (cleaning)
	제품 (product)	화장실 (toilet)
	사용(use)	효과 (effectiveness)
	뿌려 (spray)	소다 (soda)
	가격 (price)	수건 (towel)
	청소 (cleaning)	효과 (effectiveness)
	곰팡이 (mold)	사용 (use)
	화장실 (toilet)	곰팡이 (mold)
	욕실 (toilet)	제품 (product)
	곰팡 (mold)	생각 (thought)
	사용 (use)	배송 (delivery)
	뿌리 (spray)	뿌리 (spray)
	주문 (order)	지워 (erased)
	뿌려 (spray)	나오 (coming)
	최고 (best)	냄새 (odor)

	Pros	Cons
LDA	효과 (effectiveness)	곰팡이 (mold)
	냄새 (odor)	냄새 (odor)
	사용 (use)	락스 (bleach)
	뿌리 (spray)	제거 (removal)
	곰팡이 (mold)	사용 (use)
	제품 (product)	벽지 (wallpaper)
	락스 (bleach)	효과 (effectiveness)
	배송 (delivery)	배송 (delivery)
	빠른 (fast)	제품 (product)
	곰팡 (mold)	곰팡 (mold)
KcBERT + K-means	곰팡이 (mold)	락스 (bleach)
	제거 (removal)	곰팡이 (mold)
	사용 (use)	냄새 (odor)
	냄새 (odor)	사용 (use)
	효과 (effectiveness)	제거 (removal)
	뿌리 (spray)	발송 (delivery)
	청소 (cleaning)	벽지 (wallpaper)
	화장실 (bathroom)	청소 (cleaning)
	제품 (product)	화장실 (toilet)
	뿌려 (spray)	분사 (spray)
	곰팡이 (mold)	락스 (bleach)
	빠르 (fast)	곰팡이 (mold)
	사용 (use)	냄새 (odor)
	효과 (effectiveness)	효과 (effectiveness)
	배송 (delivery)	벽지 (wallpaper)
	뿌리 (spray)	사용 (use)
	제품 (product)	곰팡 (mold)
	제거 (removal)	제거 (removal)
	청소 (cleaning)	욕실 (bathroom)
	지워 (erased)	배송 (delivery)
	사용 (use)	효과 (effectiveness)
	제품 (product)	생각 (thought)
	만족 (satisfaction)	냄새 (odor)
	상품 (product)	곰팡이 (mold)
	청소 (cleaning)	분사 (spray)
	가격 (price)	분무 (spray)
	배송 (delivery)	나오 (coming)
	곰팡이 (mold)	심하 (intense)
	편리 (convenient)	제품 (product)
	성능 (performance)	지워 (erased)
	곰팡이 (mold)	냄새 (odor)
	제거 (removal)	곰팡이 (mold)
	빠르 (fast)	그래요 (be)
	배송 (delivery)	지워짐 (erased)
	효과 (effectiveness)	제거 (removal)
	사용 (use)	효과 (effectiveness)
	만족 (satisfaction)	리뷰 (review)
	최고 (best)	욕실 (bathroom)
	곰팡 (mold)	벽지 (wallpaper)
	지워 (erased)	닦이 (cleaner)

	Pros	Cons
KcBERT + K-means	곰팡이 (mold) 제거 (removal) 효과 (effectiveness) 곰팡 (mold) 청소 (cleaning) 사용 (use) 화장실 (toilet) 배송 (delivery) 냄새 (odor)	사용 (use) 곰팡이 (mold) 제품 (product) 효과 (effectiveness) 배송 (delivery) 바르 (cover) 똑같 (same) 지워 (erased) 락스 (bleach)
	욕실 (bathroom)	뿌리 (spray)
	빠른 (fast) 배송 (delivery) 청소 (cleaning) 화장실 (toilet) 욕실 (bathroom) 곰팡이 (mold) 만족 (satisfaction) 사용 (use) 제품 (product) 냄새 (odor)	
KcBERT + DCN	곰팡이 (mold) 제거 (removal) 효과 (effectiveness) 사용 (use) 배송 (delivery) 빠르 (fast) 뿌리 (spray) 지워 (erased) 청소 (cleaning) 뿌려 (spray)	효과 (effectiveness) 사용 (use) 곰팡이 (mold) 제품 (product) 생각 (thought) 배송 (delivery) 냄새 (odor) 분사 (spray) 지워 (erased) 불편 (inconvenience)
	곰팡이 (mold) 제거 (removal) 사용 (use) 효과 (effectiveness) 냄새 (odor) 청소 (cleaning) 곰팡 (mold) 뿌리 (spray) 화장실 (toilet) 제품 (product)	곰팡이 (mold) 배송 (delivery) 효과 (effectiveness) 제거 (removal) 냄새 (odor) 락스 (bleach) 분무기 (spray) 박스 (box) 사용 (use) 내용물 (contents)
	배송 (delivery) 빠른 (fast) 청소 (cleaning) 사용 (use) 화장실 (toilet) 제품 (product) 만족 (satisfaction) 곰팡이 (mold) 욕실 (bathroom) 상품 (product)	곰팡이 (mold) 효과 (effectiveness) 제거 (removal) 냄새 (odor) 벽지 (wallpaper) 심하 (intense) 사용 (use) 바르 (cover) 뿌리 (spray) 락스 (bleach)

	Pros	Cons
KcBERT + DCN	빠르 (fast) 배송 (delivery) 제품 (product) 상품 (product) 빠른 (fast) 빨라 (fast) 물건 (item) 편하 (convenient) 도착 (arrive) 와서 (arrive)	락스 (bleach) 냄새 (odor) 곰팡이 (mold) 사용 (use) 벽지 (wallpaper) 제거 (removal) 제품 (product) 발송 (delivery) 살림 (housekeeping) 뿌리 (spray)
KcBERT + UMAP + HDBSCAN (BERTopic)	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)

	Pros	Cons
KcBERT + UMAP + HDBSCAN (BERTopic)		사용 (use) 곰팡이 (mold) 제품 (product) 뿌리 (spray) 바르 (cover) 효과 (effectiveness) 지워 (erased) 똑같 (same) 힘들 (difficult) 구입 (purchase)

References

1. Anoop, V.S.; Asharaf, S. A topic modeling guided approach for semantic knowledge discovery in e-commerce. *Int. J. Interact. Multimed. Artif. Intell.* **2017**, *4*, 40–47. [CrossRef]
2. Kyeong, J.; Jiyoon, K.; Lee, J.W.; Lee, Y.; Lee, S.B. Text Mining Analysis of Consumer Perception of Food Distribution Platforms: Focusing on Topic Modeling. *J. Foodserv. Manag.* **2021**, *24*, 71–100.
3. Bumjun, L.; Heekyung, N. Food tourism market segmentation approach using topic modeling analysis: Focusing on benefits sought. *Korean J. Hosp. Tour.* **2020**, *29*, 187–204.
4. Soyeon, L.; Yeongok, K. Analysis of Apartment Interior Trend Using Topic Modeling: Focusing on 'Today's House' Review Data. In Proceedings of the KMIS 2022: 14th International Conference on Knowledge Management and Information Systems, Valletta, Malta, 24–26 October 2022; pp. 141–149.
5. Cho, M.-K.; Lee, B.-J. Comparison of service quality of full service carriers in Korea using topic modeling: Based on reviews from TripAdvisor. *J. Hosp. Tour. Stud.* **2021**, *23*, 152–165.
6. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
8. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
9. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G., Eds.; PAKDD 2013. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7819. [CrossRef]
10. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 226–231.
11. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 478–487.
12. Yang, B.; Fu, X.; Sidiropoulos, N.D.; Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–7 August 2017; pp. 3861–3870.
13. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
14. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
15. Uncovska, M.; Freitag, B.; Meister, S.; Fehring, L. Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. *NPJ Digit. Med.* **2023**, *6*, 115. [CrossRef] [PubMed]
16. Li, W.; Yin, J.; Chen, H. Supervised Topic Modeling Using Hierarchical Dirichlet Process-Based Inverse Regression: Experiments on E-Commerce Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 1192–1205. [CrossRef]
17. Alhaj, F.; Al-Haj, A.; Sharieh, A.; Jabri, R. Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 854–860. [CrossRef]
18. Ji, Y.; Ma, Y. The robust maximum expert consensus model with risk aversion. *Inf. Fusion* **2023**, *99*, 101866. [CrossRef]
19. de Groot, M.; Aliannejadi, M.; Haas, M.R. Experiments on generalizability of BERTopic on multi-domain short text. *arXiv* **2022**, arXiv:2212.08459.
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Duchesnay, É. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
21. Terragni, S.; Fersini, E.; Galuzzi, B.G.; Tropeano, P.; Candelieri, A. OCTIS: Comparing and optimizing topic models is simple! In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online, 19–23 April 2021; pp. 263–270.
22. Lee, J. KcBERT: Korean Comments BERT. In Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology, Boulder, CO, USA, 29–31 March 2019; pp. 437–440.

23. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
24. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
25. Harrando, I.; Lisena, P.; Troncy, R. Apples to Apples: A Systematic Evaluation of Topic Models. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1–3 September 2021; INCOMA Ltd.: Shoumen, Bulgaria; pp. 483–493.
26. Fang, A.; Macdonald, C.; Ounis, I.; Habel, P. Using word embedding to evaluate the coherence of topics from twitter data. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 1057–1060.
27. Available online: https://fasttext.cc/docs/en/crawl-vectors.html (accessed on 8 July 2023).
28. Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proc. GSCL* **2009**, *30*, 31–40.
29. Dieng, A.B.; Ruiz, F.J.; Blei, D.M. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [CrossRef]