Automated Keyword Filtering in Latent Dirichlet Allocation for Identifying Product Attributes From Online Reviews

Junegak Joung
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801;
Department of Industrial Engineering,
Ulsan National Institute of Science and Technology,
Ulsan 44919, South Korea
e-mails: junegak@illinois.edu; june30@unist.ac.kr

Harrison M. Kim[1]
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801
e-mail: hmkim@illinois.edu

Identifying product attributes from the perspective of a customer is essential to measure the satisfaction, importance, and Kano category of each product attribute for product design. This article proposes automated keyword filtering to identify product attributes from online customer reviews based on latent Dirichlet allocation. The preprocessing for latent Dirichlet allocation is important because it affects the results of topic modeling; however, previous research performed latent Dirichlet allocation either without removing noise keywords or by manually eliminating them. The proposed method improves the preprocessing for latent Dirichlet allocation by conducting automated filtering to remove the noise keywords that are not related to the product. A case study of Android smartphones is performed to validate the proposed method. The performance of the latent Dirichlet allocation by the proposed method is compared to that of a previous method, and according to the latent Dirichlet allocation results, the former exhibits a higher performance than the latter. [DOI: 10.1115/1.4048960]

Keyword: design automation

1 Introduction

Identifying customer needs for product design is significant because customer-driven products are the key to success in the market [1]. In this context, various studies to identify customer needs from publicly available online reviews have been conducted in the product design literature. Unsatisfactory and satisfactory product attributes were identified using the sentiment analysis [2–7]. A Kano model analysis of product attributes was performed [2,7]. The importance of product attributes was estimated to determine the priority of the product development [8–10].

To conduct the analysis of customer needs from online reviews, identifying the product attributes that customers frequently mention or evaluate is essential. Latent Dirichlet allocation (LDA), association rule mining, frequency, and word embedding have been used to identify product attributes from online reviews. LDA is a topic modeling method that is commonly used to identify topics in various domains [3]. The preprocessing for LDA is important because it affects the results of topic modeling [11]. The preprocessing of LDA in previous studies has been performed either without filtering out words not related to product features or manually. The purpose of this article is to propose automated keyword filtering in the preprocessing for LDA for producing better LDA results.

2 Related Work

When identifying product attributes from online reviews, product attributes and words corresponding to each attribute are predetermined or unknown. When they are unknown, various methods have been conducted based on association rule mining, frequency, word embedding, and LDA. Association rule mining has been used to find words that frequently occur together, and product attributes and their words have been identified by pruning rules [6,12]. The frequency-based method has been employed considering high-frequency nouns as a product attribute [8,9]. However, these methods cannot remove words that are not related to product features [10]. Word embedding-based clustering of noun words and filtering using product manuals was performed to identify product attributes [10]. Compared with previous methods, this method can effectively acquire words related to product features by filtering, but it cannot consider noun phrases as product attributes. This research considers both nouns and noun phrases and identifies product attributes from online reviews based on LDA, which is commonly used in various fields as a topic model.

Previous research of the preprocessing for LDA for identifying product attributes has been conducted (Table 1). This preprocessing includes common steps, such as lower casing; removing punctuations, stop words, and very frequently and very rarely used words; and lemmatizing or stemming [13,14]. Based on the aforementioned common steps, previous studies considered different parts of speech (POS) of words or removed words unrelated to the product. Adjectives, nouns, adverbs, and verbs were considered as POS in LDA, and each attribute was identified by interpreting its related adjective, noun, adverb, and verb based on their co-occurrence from online reviews [5,7]. Keywords of nouns and noun phrases were considered as POS in LDA by removing sentiment words and degree words. In comparison, noise keywords were either not eliminated [2] or were manually filtered out by considering the hardware or software features of the product before the LDA [3,4].

However, previous studies that considered adjectives and adverbs could not easily identify product attributes because these words are modifiers of nouns or verbs, such as "good," "great," and "quickly." Noise keywords in the LDA results make it difficult to understand product attributes. Manual efforts to remove noise

Table 1 Summary of the preprocessing for LDA in previous research

Literature	POS	Filtering method	Data
El Dehaibi et al. [3]	Adjective, noun, adverb, verb	Manual operation	French Press coffee makers
Wang et al. [5]	Adjective, noun, adverb, verb	None	Wireless mouses
Zhou et al. [7]	Adjective, noun, adverb, verb	None	Amazon product ecosystem
Bi et al. [2]	Noun	None	Smart phones
Jeong et al. [4]	Noun, noun phrase	Manual operation	Samsung Galaxy Note 5

[1]Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the JOURNAL OF MECHANICAL DESIGN. Manuscript received June 10, 2020; final manuscript received October 9, 2020; published online February 9, 2021. Assoc. Editor: Scott Ferguson.

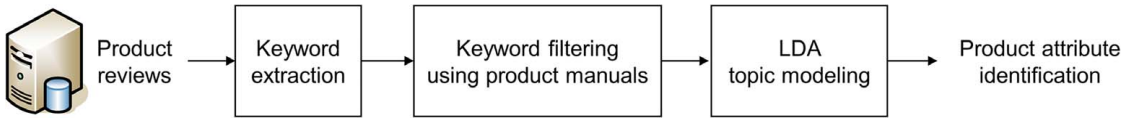

Fig. 1 Overall process

keywords are also time consuming. Therefore, this research improves the preprocessing for LDA to identify product attributes from online reviews by introducing automated filtering to remove the keywords that are not related to the product.

3 Method

The overall process to identify product attributes from online customer reviews is as follows (Fig. 1). The inputs of the method are product reviews from customers, and the outputs are the product attributes that customers frequently mention and evaluate in their reviews. After collecting customer reviews of a target product, first, keywords are extracted from each review using keyword extraction tools. Second, the noise keywords that are not related to the product features are filtered out using product manuals. Finally, the product attributes are identified by LDA based on product-related keywords that customers mention together. The keyword extraction and keyword filtering steps are automated, whereas the interpretation of the LDA results requires human involvement.

3.1 Data Collection and Keyword Extraction. Customer reviews of a target product are collected from websites, such as Amazon, eBay, and Best Buy. Web scraping assists in automatically collecting information, such as title, review, date, rating, and user name, from the html documents of web pages. To refine the review data for the analysis, duplicated reviews that appear more than once are removed by identifying reviews with the same titles and contents. Emojis, emoticons, and newline characters in each review are stripped by identifying specific patterns, such as "U+1F600," "U+1F603," "U+1F604," and ":D."

After collecting the reviews, keywords of nouns and noun phrases can be extracted using open-source keyword extraction tools, such as rapid automatic keyword extraction or IBM Watson Natural Language Understanding (NLU). Here, Watson NLU is used to automatically extract keywords by removing stop words, such as "he," "she," "about," and "that." Watson NLU can reduce the time for keyword extraction and ensure identification of noun phrases that are relatively difficult to extract compared to nouns, because it applies large-scale data and deep learning methods [4]. Subsequently, text preprocessing is conducted as follows [4,13]: uppercases are transformed to lowercases (e.g., "Battery" is converted to "battery"), punctuations are removed (e.g., "4g-lte" is converted to "4g lte"), words are lemmatized (e.g., "images" is converted to its root form "image"), and words that occur either very frequently or very rarely are eliminated. However, Watson NLU cannot consider a noun in a noun phrase, such as "adjective + noun" and "noun + noun" (Fig. 2(*a*)). POS tagging is used to automatically extract nouns from each noun phrase. Consequently, each review is structured into keywords that contain nouns and noun phrases (Fig. 2(*b*)).

3.2 Keyword Filtering Using Product Manuals. Keyword filtering is proposed to eliminate numerous noise keywords that are not related to the product features using product manuals. Product manuals are called as user manuals and are written from the perspective of the user, including significant amount of customer terminology. Product manuals contain numerous keywords related to the product attributes for introducing a product to users. The product manual has been adopted to identify relevant keywords

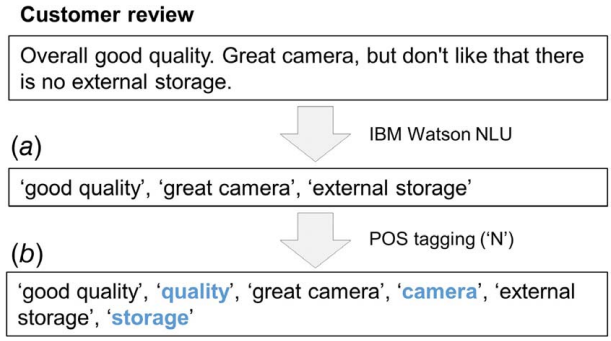

Fig. 2 Flowchart of keyword extraction

of product usage and attributes. The strategy for keyword filtering is to remove the noise keywords mentioned in customer reviews if they do not occur commonly in product manuals [10]. The keyword filtering algorithm proceeds in three steps:

Step 1: A set of product manual documents of the target product is collected, and subsequently keywords are extracted from them, similar to extracting keywords from customer reviews. The extracted keywords from product manual documents are nouns and noun phrases, and the preprocessing of lower casing, removing punctuation, lemmatizing, and extracting nouns from each noun phrase is next conducted. This preprocessing shares numerous processes with the preprocessing of reviews, thereby facilitating comparisons of the keywords in both types of documents. Different from the preprocessing of reviews, very frequently or very rarely occurring keywords in the product manuals are not removed, owing to the difference between the number of manual documents and the number of reviews. These keywords can be considered as product-related keywords in a small number of product manual documents.

Step 2: The proportion of a keyword that is identified in customer reviews is calculated based on the keyword frequency from each manual document. $P(k_i)$ is computed as the frequency of keyword k_i divided by the total number of keywords in the manual document. For example, if five manual documents are used, five $P(k_i)$ are calculated in each document.

Step 3: A one-sample t-test is conducted with the average proportion of a keyword.

$$H_0: Avg(P(k_i)) = 0$$
$$H_1: Avg(P(k_i)) > 0 \quad (1)$$

If the null hypothesis (H_0) is not rejected, the average proportion of keyword k_i is statistically equal to 0. Keyword k_i that does not occur in product manuals is filtered out. Alternatively, keyword k_i is regarded as a product-related keyword because it is common to product manual documents. A previous study manually removed noise keywords in the preprocessing for LDA. However, the proposed method automates these manual operations by introducing the keyword filtering algorithm using keywords from product manuals. Consequently, a matrix of reviews and product-related keywords is prepared as the input of the LDA.

3.3 Topic Modeling Using Latent Dirichlet Allocation. LDA is used to identify product attributes using a review-keyword

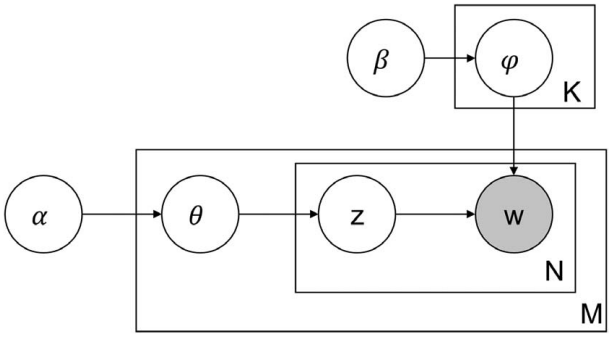

Fig. 3 Graphical model representation of LDA

matrix. The output of LDA is a topic-keyword matrix. Each topic is named by interpreting the logical relation between the top keywords in the topic and the corresponding relative probabilities [5,15]. The label of each topic can be considered as an attribute of the target product [4,5,7].

LDA is a powerful probabilistic topic model that summarizes a large amount of textual data by identifying hidden topics [16]. The LDA model assumes that each review document is considered as a mixture of a set of topic probabilities, and each topic is considered as a mixture of an underlying set of words. The graphical model of LDA is described in Fig. 3. The definitions of the modeling notation are introduced as follows:

M = number of customer reviews
N = number of words in a review
K = number of topics
α = parameter of the Dirichlet prior on the per-review topic distribution
β = parameter of the Dirichlet prior on the per-topic word distribution
θ_i = topic distribution for review i (the sum of θ_i is 1)
φ_k = word distribution for topic k
z_{ij} = topic for the j^{th} word in review i
w = specific word

Based on these notations, LDA involves a generative process as follows:

Step 1: Choose $\theta_i \backsim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$
Step 2: Choose $\varphi_k \backsim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$
Step 3: For each word position i, j, where $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N_i\}$
Choose a topic $z_{ij} \backsim$ Multinomial (θ_i)
Choose a word $w \backsim$ Multinomial $(\varphi_{z_{ij}})$

Based on the aforementioned process, parameters such as α, β, θ_i, φ_k, and K, for executing LDA, need to be estimated. However, the exact parameter estimation of the LDA model is intractable; therefore, approximate estimation methods are used. Variation expectation maximization algorithm [16], Gibbs sampling methods [17], and collapsed variational Bayes approximation [18] can be used to infer the parameters such as α, β, θ_i, and φ_k. Perplexity measure [16], average similarity between all the topics, and topic coherence [19] can be used to determine K. An LDA model with a low perplexity value and average similarity between various K values represents the model best. An LDA model with a high topic coherence indicates the model best. The optimal number of topics K may be determined from similar evaluation levels in addition to the number of topics K with the best value. Topic coherence is applied, and it determines optimal number of topics K to achieve a high correlation with human ratings [20]. Topic coherence assumes that if a topic is more interpretable, the top pairs of words related to the topic will co-occur more frequently in the reviews. For example, a topic with top words "screen" and "display" is more interpretable or coherent if both the words are mentioned together in numerous customer reviews. Topic coherence of an LDA model T consisting of k topics is calculated as follows:

$$Coh(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} sim(w_i, w_j)$$
$$Coh(T) = \frac{1}{k} \sum_{h=1}^{k} Coh(t_h) \tag{2}$$

4 Case Study

A case study of the Android smartphones of a manufacturer was conducted to demonstrate the proposed method. This case study was chosen because it is a popular model and is likely to generate many reviews. Identifying product attributes from these reviews is highly likely to represent the population. A web scraper chrome extension (e.g., WebScraper.io) was employed to collect the customer reviews of verified purchases in the cell phone category of Amazon.com. Figure 4 shows an example of the review data. Customer reviews are assumed to be authentic, because consumers write reviews of verified purchase voluntarily. After removing the overlapped reviews with the same titles and contents and stripping emojis, emoticons, and newline characters from each review, 33,779 reviews of Android smartphones were obtained from April 2014 to September 2019. The Android smartphones included three series models of similar sizes. These smartphones share numerous common features with a smartphone released by a specific manufacturer. Emojis and emoticons were removed because this research uses the keywords of each review to identify product attributes. The original form of reviews without eliminating emojis and emoticons can be used for further analysis if the sentiment analysis containing emojis and emoticons is conducted as the subsequent analysis.

4.1 Extracting Keywords. From the collected reviews, 51,011 keywords of nouns and noun phrases were extracted using Watson NLU of PYTHON. This was proceeded by the text

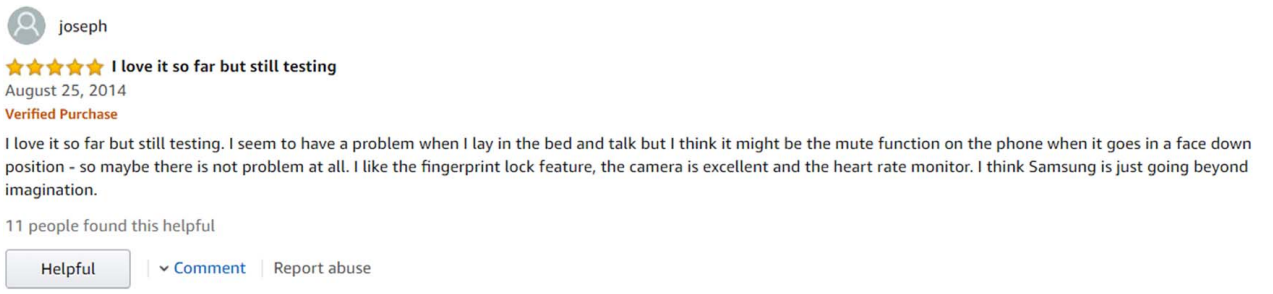
joseph
I love it so far but still testing
August 25, 2014
Verified Purchase
I love it so far but still testing. I seem to have a problem when I lay in the bed and talk but I think it might be the mute function on the phone when it goes in a face down position - so maybe there is not problem at all. I like the fingerprint lock feature, the camera is excellent and the heart rate monitor. I think Samsung is just going beyond imagination.
11 people found this helpful
Helpful
Comment
Report abuse

Fig. 4 Example of the collected reviews

Table 2 Keyword statistics after preprocessing review data

Step	Number of nouns	Number of noun phrases	Total
Keyword extraction	10,880	40,131	51,011
Text preprocessing	4860	7123	11,983
Keyword filtering	940	286	1226

preprocessing. From the entire review documents, one very frequently occurring keyword (e.g., "phone") that was present in more than 50% of the documents and very rarely used 39,027 keywords (e.g., "manufacture sticker," "fake cellphone," and "noodle") appearing in only one document were removed. These keywords were eliminated because having numerous common keywords is not relevant in LDA results because they are mentioned very frequently with other keywords, and in addition, local keywords do not affect LDA results [4]. The case study was selected as a specific smartphone series, so the brand and model names were also removed. Nouns from a noun phrase were extracted by POS tagging. The preprocessing was executed using the natural language toolkit package of PYTHON. Each of the 33,779 reviews were structured into keywords, which included 4860 nouns and 7,123 noun phrases (Table 2).

4.2 Filtering out Noise Keywords. After the text preprocessing, most of the keywords were removed; however, there were still noise keywords, such as "star, "thing," and " amazon." The keyword filtering algorithm was implemented using product manuals. First, three manual documents of the smartphone models with the most reviews in each smartphone series were collected, because the manual documents of the same smartphone series were similar. Subsequently, keywords of nouns and noun phrases in each manual document were extracted and refined according to Sec. 3.2 (Table 3). Second, 11,983 $P(K_i)$ of the keywords identified in the reviews from each manual document were calculated. Finally, one-sample t-tests were performed at various confidence levels of 70%, 80%, 90%, and 95%. Product-related keywords were identified as 1226, 791, 647, and 425 in order, because the null hypothesis was strictly rejected as the confidence level became higher. A confidence level of 70% was chosen to identify product-related keywords maximally because considering numerous product-related words can lead to topic modeling from many reviews including these keywords. Consequently, 1226 product-related keywords such as "screen," "camera," "app," "call," "internet," and "battery" were automatically identified (Table 2).

4.3 Identifying Product Attributes. After identifying 1226 product-related keywords, 8726 reviews that did not contain these keywords were considered as noise reviews and excluded from the LDA analysis. These reviews were typically short and unsuitable for identifying product attributes because customers did not mention the hardware or software features of the product as follows:

R1: "Good phone. Bought for my wife. She loves this phone!"
R2: "Two Stars. Not what I expected it"
R3: "Five Stars. I love it, thanks"

For LDA, a $25{,}053 \cdot 1226$ review-keyword matrix was formed.

Table 3 Keyword statistics based on product manuals

Product manual	Number of nouns	Number of noun phrases	Total
A manual	813	1447	2260
B manual	881	1533	2414
C manual	1100	2524	3624

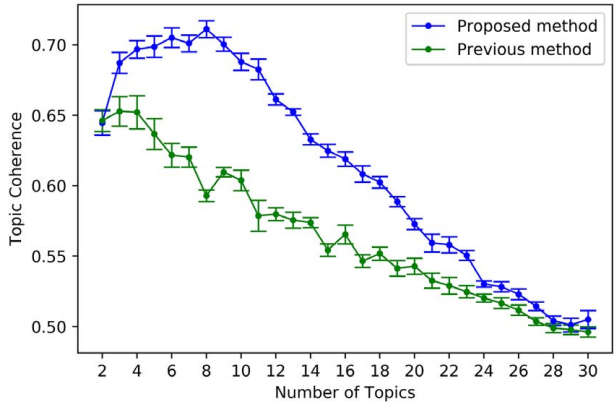

Fig. 5 Comparison of topic coherence (in mean ± standard error) from topics 2 to topics 30

By taking the input matrix, the Gensim library [21] of PYTHON, which applies the variation expectation maximization algorithm, was used to infer the parameters, such as α, β, θ_i, and φ_k, in LDA. The topic coherence was estimated to avoid the overclustering problem and determine the optimal number of topics K. Various measures of topic coherence, such as "C_uci," "C_npmi," "C_umass," and "C_v," can be considered when calculating and combining similarities between words in a topic (Eq. (2)). "C_uci" and "C_npmi" are estimated based on pointwise mutual information and normalized pointwise mutual information, respectively, and "C_umass" is calculated based on document frequencies. "C_v" is measured based on a one-set segmentation of the top words and a confirmation measure that uses normalized pointwise mutual information and the cosine similarity. The "C_v" measure was applied because the measure presents a larger correlation with human ratings than other measures [22]. The perplexity value and the average similarity between all the topics were calculated; however, they were not appropriate for the keyword filtering of the proposed method. The perplexity values continued to decrease as K increased, and the average similarity between all the topics was the lowest when K was 2. Therefore, the optimal model with the best value could not be found. A tenfold cross-validation strategy was used to identify the maximum topic coherence (Fig. 5). The number of topics selected were 8 based on the maximum topic coherence value (0.711 ± 0.006).

Each topic was named by identifying the logical relationship between their top-30 keywords and their corresponding relative probabilities (Table 4). The top 30 keywords were considered by sufficiently providing the most relevant terms in each topic although top 10 and top 20 keywords could be included [23]. Keywords related to each topic were identified from the top 30 keywords, and frequent keywords were arranged in a descending order according to the probability that is relevant to the topic. The ratio indicates the percentage of reviews that is most relevant to each topic in all the reviews [23]. Most topics were easy to identify without investigating the review comments. For example, the second topic was named "screen" attribute considering its top related keywords (e.g., "screen," "case," "size," "display," "protector," and "glass"). The third, fourth, fifth, sixth, seventh, and eighth topics were also named. However, the first topic, "product check" attribute, was named by examining the typical reviews, which included the most probable product-related keywords, such as "product," "box," "condition," and "model." The typical reviews contained "I was afraid about this, but now, nothing at all, excellent product and perfect conditions. Just to say about delivery, its slow. U have to wait." This attribute was not directly related to the hardware or software of the Android smartphones; however, it was identified because it represented the initial condition or quality of the ordered product. "Product check" was 26.7% most frequently mentioned attribute by the customers. Among the remaining seven attributes, the customers were more

Table 4 Eight product attributes from the android smartphones

Number	Product attribute	Frequent keywords	Number of words	Ratio (%)
1	Product check	Product, problem, seller, box, device, condition, version, model, warranty, item, description, replacement, support, return	14	26.7
2	Screen	Screen, case, size, display, protector, glass, cover, screen protector, pocket, touch	10	16.9
3	Camera	Camera, quality, picture, video, photo, light, front, picture, resolution, image	10	12.9
4	App	Apps, android, update, app, notification, email, application, mail, file	9	12.5
5	Communication	Call, network, data, text, message, lte, internet, signal, voice, connection, contact, fi, gps	13	9.7
6	Battery	Battery, life, battery life, charge, use, power, drain, battery drain, fast charging, battery charge, battery power	11	7.9
7	Card slot	Card, sim, sim card, sd, slot, sd card, dual sim, memory card, pin, microsd	10	7.5
8	Accessory	Charger, port, cable, accessory, plug, usb, earphone, wall, jack, microphone, assistant, wireless charging	12	5.8

Table 5 Top five product-related keywords of "product check," "screen," "camera," and "app" attributes

Product check	Sentiment intensity	Screen	Sentiment intensity	Camera	Sentiment intensity	App	Sentiment intensity
Service	−0.054	Screen	−0.097	Camera	0.433	Apps	−0.201
Condition	0.635	Case	−0.140	Picture	0.357	Android	0.130
Box	−0.097	Size	0.499	Video	0.130	App	−0.352
Warranty	−0.494	Button	−0.514	Photo	0.336	Setting	−0.198
Delivery	0.414	Fingerprint	−0.059	Resolution	0.428	Music	0.130

concerned about "screen," "camera," and "app." These attributes were mentioned more than 10% in all the topics.

4.4 Validation of Keyword Filtering. The proposed keyword filtering was validated because it yielded better LDA results than a previous study. The topic coherence was used as a performance measure of the LDA topic modeling. Better preprocessing for LDA indicates a higher topic coherence over numerous topics. In the previous studies, text preprocessing of lower casing, lemmatizing, and removing punctuation and keywords that occur either very frequently or very rarely without removing the noise keywords was conducted [2,5,7], and keyword filtering was manually performed [3,4]. The proposed method was compared to the case in which the text preprocessing of Sec. 4.1 was performed without eliminating the noise keywords in Sec. 4.2 because the manual removal of noise keywords depends on subjective judgment. The topic coherence values of the LDA models by the proposed method were higher than those obtained by the previous method over numerous topics (Fig. 5). By the previous method, the optimal number of topics was three, and the maximum topic coherence value was 0.653, which was lower than the value achieved by the proposed method (0.711). Furthermore, the previous method could not easily identify the product attributes in the LDA results owing to the top noise keywords, such as "star," "verizon," "sprint," "amazon," and "thing." In comparison, the proposed method could easily identify the product attributes because of the top product-related keywords, such as "screen," "display," "camera," "call," "network," "battery," "sim card," and "charger" (Table 4).

5 Discussion

This section discusses the application of the proposed method for product design and the one-sample t-test for the keyword filtering.

5.1 Application of Proposed Method for Product Design. The proposed method can be the basis for product designers to identify the customer preferences of both product attributes and their specific product features. The sentiments for each product attribute and those for its keywords in this study can be measured using a machine-learning-based sentiment classifier. Sentiment analysis can provide information on which product attributes customers are satisfied and dissatisfied with and which features are satisfied and dissatisfied with these attributes. In the case study, the keyword sentiment analysis of IBM Watson was used to measure the sentiment intensities of the keywords, which ranged from −1 (i.e., negative) to 1 (i.e., positive). The sentiment of each keyword was calculated by averaging their sentiment intensities in the overall review, and the sentiment of each product attribute was measured by averaging the sentiments of these keywords. The sentiment intensities of eight product attributes as well as the top five keywords in each attribute could be measured; however, "product check," "screen," "camera," and "app" of 10% ratio or more were derived (Table 5). The attributes were 0.018, −0.074, 0.363, and −0.128 in order. "Product check" and "screen" attributes were close to 0; therefore, there were no clear positive or negative responses from the customers: the "camera" attribute presented positive responses, whereas the "app" attribute showed negative responses. Under the "product check" attribute, "condition" and "delivery" were evaluated positively, whereas "warranty" was evaluated negatively. Under the "screen" attribute, "size" was better, whereas "button" was badly evaluated. Under the "camera" attribute, the top keywords were rated as good overall. In the "app" attribute, "android" and "music" were evaluated positively, whereas "app" and "setting" were evaluated negatively. Thus, the sentiment analysis at the attribute and feature levels based on this study will be useful for product designers to identify customer needs.

5.2 Use of One-Sample T-Test for Keyword Filtering. In this research, a one-sample t-test is conducted to investigate which keywords commonly occur in product manuals, and the use of the one-sample t-test is examined. The one-sample t-test assumes that the samples follow a normal distribution. If there are more than 30 samples, the normality can be assumed by the central limit theorem [24]. Even if the sample size is small, the

normality can be assumed if a case study selects similar product models with similar manual documents. In the case study, three samples from three manual documents were used in the one-sample t-test for the keyword filtering, and it was examined whether these samples had a normal distribution. Three samples were too few to confirm this assumption; therefore, the manual document of each series model was added. The normality test was performed by calculating six $P(k_i)$ of each keyword from six product manuals using 11,983 keywords before keyword filtering. The Shapiro–Wilk W test [25] was used to check that a total of six samples followed a normal distribution at a 99 % confidence level. Approximately 93.7% (11225/11983) of the keywords followed a normal distribution. If the manual documents of each series model are added, the ratio of the keywords with the normality will be higher. Therefore, the use of the one-sample t-test for keyword filtering is appropriate in the case study. If the difference between the product manuals is large, numerous product manuals may be collected to ensure the normality; alternatively, a Wilcoxon signed-rank test may be conducted instead of the normality assumption.

6 Conclusion and Future Work

This article proposes keyword filtering in LDA to identify product attributes from online customer reviews. The proposed method improves the automation in keyword preprocessing compared to previous LDA applications. The case study of Android smartphones demonstrates that the proposed method yields better LDA results to identify product attributes than a previous method.

Future studies can be tested on more data with the proposed method, and it can apply sentiment analysis and deep learning to identify the satisfaction, importance, and Kano category of the identified product attributes. The proposed keyword filtering may include some noise keywords depending on the product manuals and the confidence level of the one-sample t-test. Future research can improve this by robust filtering. The proposed method identifies product attributes by focusing on product features frequently mentioned by customers in online reviews. Future studies may attempt to find rarely mentioned product features and their responses from customers using keywords that do not appear in product manuals. The keyword filtering may be unsuitable for the case with no product manuals or product manuals that are very different from customer terminology. Future studies can apply word embedding to solve this terminology problem. For example, synonyms of "camera" and "app," such as "cam" and "application," can be automatically considered based on word embedding [26].

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1I1A1A01063298).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The authors attest that all data for this study are included in the paper. No data, models, or code were generated or used for this paper.

Nomenclature

t = t top-ranked words
H_0 = null hypothesis
H_1 = alternative hypothesis
$P(k_i)$ = proportion of the keyword i in the manual document
$Coh(t_h)$ = coherence of a single topic t_h
$sim(w_i, w_j)$ = similarity of word i and word j
$Coh(T)$ = overall coherence of a LDA model T consisting k topics

References

[1] Chen, W., Conner, C., and Yannou, B., 2015, "User Needs and Preferences in Engineering Design," ASME J. Mech. Des., **137**(7), p. 068001.
[2] Bi, J.-W., Liu, Y., Fan, Z.-P., and Cambria, E., 2019, "Modelling Customer Satisfaction From Online Reviews Using Ensemble Neural Network and Effect-Based Kano Model," Int. J. Prod. Res., **57**(22), pp. 7068–7088.
[3] El Dehaibi, N., Goodman, N. D., and MacDonald, E. F., 2019, "Extracting Customer Perceptions of Product Sustainability From Online Reviews," ASME J. Mech. Des., **141**(12), p. 121103.
[4] Jeong, B., Yoon, J., and Lee, J.-m., 2017, "Social Media Mining for Product Planning: A Product Opportunity Mining Approach Based on Topic Modeling and Sentiment Analysis," Int. J. Inform. Manag., **48**, pp. 280–290.
[5] Wang, W., Feng, Y., and Dai, W., 2018, "Topic Analysis of Online Reviews for Two Competitive Products Using Latent Dirichlet Allocation," Electron. Commerce Res. Appl., **29**, pp. 142–156.
[6] Zhou, F., Jiao, R. J., and Linsey, J. S., 2015, "Latent Customer Needs Elicitation by Use Case Analogical Reasoning From Sentiment Analysis of Online Product Reviews," ASME J. Mech. Des., **137**(7), p. 071401.
[7] Zhou, F., Ayoub, J., Xu, Q., and Jessie Yang, X., 2020, "A Machine Learning Approach to Customer Needs Analysis for Product Ecosystems," ASME J. Mech. Des., **142**(1), p. 011101.
[8] Jiang, H., Kwong, C., and Yung, K., 2017, "Predicting Future Importance of Product Features Based on Online Customer Reviews," ASME J. Mech. Des., **139**(11), p. 111413.
[9] Rai, R., 2012, "Identifying Key Product Attributes and Their Importance Levels From Online Customer Reviews," ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering, Chicago, IL, Aug. 12–15, pp. 533–540.
[10] Suryadi, D., and Kim, H., 2018, "A Systematic Methodology Based on Word Embedding for Identifying the Relation Between Online Customer Reviews and Sales Rank," ASME J. Mech. Des., **140**(12), p. 121403.
[11] Denny, M. J., and Spirling, A., 2018, "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It," Political Anal., **26**(2), pp. 168–189.
[12] Hu, M., and Liu, B., 2004, "Mining and Summarizing Customer Reviews," Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, Aug. 22–25, pp. 168–177.
[13] Boyd-Graber, J., Mimno, D., and Newman, D., 2014, Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements, E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg, eds., Vol. 225255, CRC Press, Boca Raton, FL.
[14] Mankad, S., Han, H. S., Goh, J., and Gavirneni, S., 2016, "Understanding Online Hotel Reviews Through Automated Text Analysis," Service Sci., **8**(2), pp. 124–138.
[15] Guo, Y., Barnes, S. J., and Jia, Q., 2017, "Mining Meaning From Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation," Tourism Manage., **59**, pp. 467–483.
[16] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**, pp. 993–1022.
[17] Griffiths, T. L., and Steyvers, M., 2004, "Finding Scientific Topics," Proc. Natl. Acad. Sci. USA, **101**(Suppl 1), pp. 5228–5235.
[18] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W., 2009, "On Smoothing and Inference for Topic Models," Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Quebec, Canada, June, pp. 27–34.
[19] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., 2011, "Optimizing Semantic Coherence in Topic Models," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, pp. 262–272.
[20] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M., 2009, "Reading Tea Leaves: How Humans Interpret Topic Models," Advances in Neural Information Processing Systems 22, Vancouver, British Columbia, Canada, Dec. 7–10, pp. 288–296.
[21] Rehurek, R., and Sojka, P., 2010, "Software Framework for Topic Modelling With Large Corpora," Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 22.
[22] Röder, M., Both, A., and Hinneburg, A., 2015, "Exploring the Space of Topic Coherence Measures," Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, Feb. pp. 399–408.
[23] Sievert, C., and Shirley, K., 2014, "Ldavis: A Method for Visualizing and Interpreting Topics," Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, June 27, pp. 63–70.
[24] Johnson, O., 2004, *Information Theory and the Central Limit Theorem*, Imperial College Press, London, UK.
[25] Ghasemi, A., and Zahediasl, S., 2012, "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," Int. J. Endocrinol. Metabolism, **10**(2), p. 486.
[26] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., 2013, "Distributed Representations of Words and Phrases and Their Compositionality," Advances in Neural Information Processing Systems 26, Harrahs and Harveys, NV, Dec. 5–8, pp. 3111–3119.