Finding Social Networks Among Online Reviewers for Customer Segmentation

Seyoung Park
Enterprise Systems Optimization Laboratory,
Department of Industrial and Enterprise Systems Engineering,
University of Illinois Urbana-Champaign,
Urbana, IL 61801
e-mail: seyoung7@illinois.edu

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

Recently, online user-generated data have emerged as a valuable source for industrial applications. In the consumer product area, many studies analyze online data and draw implications on product design. However, most of them treat online customers as one group with the same preferences, while customer segmentation is a key strategy in conventional market analysis. This paper proposes a new methodology based on text mining and network analysis for online customer segmentation. First, the method extracts customer attributes from online review data. Then, a customer network is constructed based on these attributes and predefined networking rules. For networking, a new concept of "topic similarity" is proposed to reflect social meaning in the customer network. Finally, the network is partitioned by modularity clustering, and the resultant clusters are analyzed to understand segment properties. We validate our methodology using real-world data sets of smartphone reviews. The result shows that the proposed methodology properly reflects the heterogeneity of the online customers in the segmentation result. The practical application of customer segmentation is presented, illustrating how it can help companies design target-customer-oriented products. [DOI: 10.1115/1.4055624]

Keywords: text mining, customer networks, segmentation, data-driven design, design automation

1 Introduction

In the consumer market, the segmentation, targeting, and positioning (STP) strategy [1,2] is one of the most popular strategies. Figure 1 shows the overview of the STP strategy. The basic assumption is that customers have different characteristics and preferences, so the strategy starts with segmentation. There are two types of market segmentation: product segmentation and customer segmentation. The prior divides products in the market based on their specifications, and the latter divides customers based on their attributes. The STP strategy is more focused on customer segmentation. In the first stage, a company divides a broad customer base into subgroups that share similar characteristics. In the next stage, targeting, the company evaluates each segment based on various criteria such as brand power, market size, and future demand. Then, the company determines how many and which segments to enter. In the positioning stage, the company chooses a frame of reference that identifies the target market and relevant competition, and then locates the product in the frame. For example, the company can draw a positioning map based on two criteria: (i) whether the product image is modern or classic and (ii) whether the product has state-of-the-art features or focuses on easy usages. Competitors are mapped into this positioning map, and then the company selects the location of its new product. Finally, based on the selected position and target customers' needs, the company designs a new product. The design includes determining main specifications and developing selling points.

In industry, many companies have been adopting the STP model. Regarding customer segmentation, companies usually hire a market research firm and conduct surveys to obtain data. Demographics and socioeconomics are the most widely used attributes in customer segmentation [3]. These attributes include age, gender, occupation, and education level, and they can be easily obtained by properly designing survey questionnaires [4]. However, a survey has a limitation in that it requires much time and cost.

Recently, many studies have been utilizing online user-generated data in their research. In data-driven design [5], these studies analyze online data to understand customers' preferences and draw design implications. These implications include customer satisfaction for product features [6,7], usages [8], purchase behavior [9], etc. However, most of them do not consider customer segmentation. They analyze the whole customer base assuming that all of them have similar preferences. This is a significant gap between the conventional customer survey and data-driven customer analysis, i.e., the gap between the field and academia. To close this gap, this paper aims to provide a method for customer segmentation in the field of online customer analysis.

The rest of the paper is organized as follows. In Sec. 2, relevant studies will be introduced and reviewed. Section 3 will explain the details of the proposed methodology. In Sec. 4, the methodology will be conducted on actual online review data, and the result will be presented. Section 5 will evaluate the results and validate the proposed methodology. Finally, in Sec. 6, the contribution of this research will be summarized, and future works will be discussed.

2 Literature Review

In this section, two relevant topics will be discussed. The first topic is segmentation in data-driven design. The previous studies on this topic will be presented and their limitations will be discussed. The second topic is network analysis. In this topic, the application of networks will be discussed. Also, a network clustering method will be explained.

2.1 Online Customer Analysis. In 1956, Smith [10] introduced the concept of market segmentation. "Market segmentation consists of viewing a heterogeneous market (one characterized by divergent demand) as a number of smaller homogeneous markets in response to differing product preferences among important market segments." As the concept is based upon the demand side of the market, it is often referred to as customer segmentation in

[1]Corresponding author.
Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received April 14, 2022; final manuscript received September 9, 2022; published online October 10, 2022. Assoc. Editor: Christopher Hoyle.

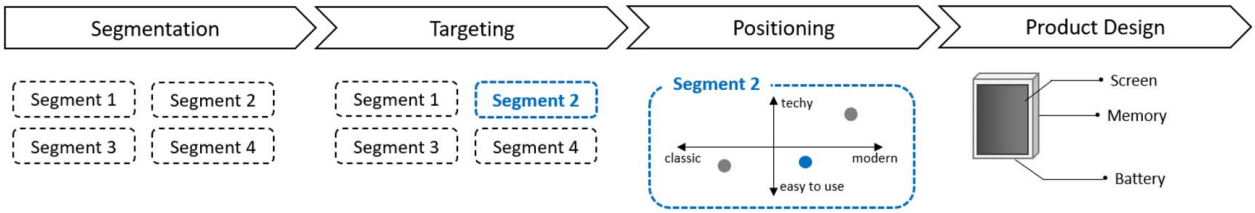

Fig. 1 STP strategy

marketing research. Kotler [1] claims two broad groups of variables to segment consumer markets: descriptive characteristics and behavioral considerations. The former includes geographic, demographic, and psychographic characteristics. The latter includes consumer responses to benefits and usage occasions. In conventional market research, these customer attributes are obtained by surveys and interviews, which takes much time and cost.

With the increasing amount of online user-generated data and the development of data analysis techniques, Big Data analytics draw significant attention. It provides valuable resources and powerful methodologies to support the data-driven decision-making process [11]. As a result, many studies have been utilizing online data to understand customers' behaviors and requirements. In the consumer product area, one of the main topics is to identify features of customer interests. Researchers proposed diverse methods for features extraction using natural language processing techniques such as latent Dirichlet allocation (LDA) [12], Word2Vec [13], and bootstrapping algorithm [7]. Another main topic is to draw design implications. Various methods have been proposed, such as extracting comparative feature importance [9,14], discovering new product features [15], and deriving strategies for features [16]. These studies draw design implications by analyzing a large set of online customer data with little cost in a short time. However, they have a limitation in that they treat the customer base as one group with the same characters. This is because customer information used in conventional segmentation is hard to obtain from online data. Nevertheless, customer segmentation is desired in online user-generated data analysis because it is a key strategy in the consumer product market.

In the previous studies about customer segmentation, cluster analysis (CA) is one of the most widely used methods [3]. Few studies in online user analysis adopted CA for customer segmentation. Wang and Chen [17] used K-means clustering to segment customers in the automobile market. However, the customer attributes used in the study are demographic characteristics collected by survey. This is not applicable for research based on user-generated online data. Park and Lee [18] conducted text mining on VoC documents, and then counted the frequency of features mentioned by each customer. The resultant VoC vector represents the customer's characteristics. The authors applied K-means clustering on the VoC vectors to segment the customers in the cellphone market. The number of clusters (K) was manually set as 10. However, 5 out of 10 resultant segments are ignored due to the small size. Moreover, the quality of the segmentation result was not discussed in the study. Suryadi and Kim [19] extracted customer attributes from online reviews for laptop products. Specifically, they analyzed the sentiment for product features and the frequency of the sentiment in each review. The resultant vector consists of product features with sentiment polarity (positive/negative). The authors conducted X-means clustering for customer segmentation, which resulted in 30 segments. The result is not applicable since it has too many segments, and the quality of resultant segments was not evaluated in the study.

This paper proposes a new methodology for online customer segmentation. The method suggests network construction and partitioning to address the limitation of vector clustering mentioned above—too many segments and lack of evaluation. The result of this study is distinct from the previous works by providing the appropriate number of segments for industrial applications and validating the quality of resultant groups with quantitative and qualitative evaluations. The details will be explained in Sec. 5.

2.2 Network Analysis. A graph is a convenient method for describing real-world situations. A graph or network is defined by an ordered triple $(V(G), E(G), \psi_G)$ consisting of a set of vertices $V(G)$, a set of edges $E(G)$, and an incidence function ψ_G that defines a pair of vertices [20]. Its applications can be found in many disciplines such as social science [21], transportation [22], and web [23]. Researches in the product design domain also utilized network analysis. Some of them conducted product segmentation by constructing product association networks. Netzer et al. [24] created a product network by analyzing user-generated online text data. They detected car models in the text data and connected the models based on the lift value, shown in Eq. (1)

$$\text{lift}(A, B) = \frac{P(A \mid B)}{P(A)} = \frac{P(A \cap B)}{P(A) \times P(B)} \tag{1}$$

The value indicates the likelihood of co-occurrence of two items in an incident. In Ref. [24], it measures how likely two vehicles are co-considered by a customer. $P(A)$ is the probability of occurrence of product A in a given message, and $P(A|B)$ is the probability of product A appearing in a message mentioning product B. When two products are independent, lift = 1. If two products are co-considered more likely than expected by chance, lift >1, and vice versa. Wang and Chen [17] analyzed choice sets of online users and also construct a product network using lift. In their study, $P(A)$ represents the probability of occurrence of product A in a choice set, and $P(A, B)$ is the probability that both A and B appear in the same choice set. Sosa et al. [25] proposed a network approach for improving the design of complex products. They analyzed the relationship among components of the product and represented the product schematic as a network of product components. The effect of design change was studied by monitoring the changes in network properties. These studies provided useful approaches for using a graph in product design. However, most of them were focused on product networks. A customer network for product design has been rarely studied. This paper proposes a new method for constructing a customer network so that segment-wise design implications can be obtained.

Once the target network is constructed, the network can be analyzed in many aspects [26]: centrality indicates the most important node or influential node within a network [27]; a clique detects a subgraph in which all nodes are connected to each other [28]; network clustering assigns a set of nodes to communities such that nodes in the same community are more similar to each other than to those in other communities [29]. In this study, network clustering was considered to divide a customer base into separate groups. There are various clustering techniques such as spectral clustering based on the Laplacian of a graph [30], the Girvan–Newman method based on the iterative elimination of edges [31], and modularity clustering based on the iterative grouping of nodes [32]. This study adopts modularity clustering for two reasons [33]: (i) it is the most used method due to its successes in many social and biological networks [34,35] and (ii) it

automatically determines the optimal number of clusters

$$Q(C) = \frac{1}{2M} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2M} \right) \delta_{ij} \tag{2}$$

Modularity is defined as Eq. (2) where d_i and d_j represent the degree of node i and j, respectively. A is an adjacency matrix. M denotes the total edge weights, and δ_{ij} is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same community} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the Louvain method [32] for modularity clustering, each node initially becomes a community. For a network with n nodes, the modularity clustering starts with n communities. The initial modularity Q is 0. Next, for every two communities, the algorithm merges them and calculate the new Q of the network. Among all possible cases, the case that leads to the biggest increase of $Q > 0$ is selected. Now, the network has $n - 1$ communities. This process of merging two communities and calculating new Q is repeated. The number of communities decreases by 1 at each iteration, and the algorithm keeps the iteration until $\Delta Q \leq 0$. The Louvain method is a greedy algorithm with the run time of $O(n \cdot \log n)$. For networks of several thousand nodes, it is among the best algorithms for modularity clustering [32].

3 Methodology

The proposed methodology consists of three stages, as shown in Fig. 2. First, customer attributes are extracted from online data. Then, a customer network is constructed based on these attributes and predefined networking rules. In the final stage, the customer network is partitioned by the modularity clustering, and the resultant segments are analyzed.

3.1 Customer Attribute Extraction. The first stage is extracting attributes of online customers. As mentioned in Sec. 1, the conventional methods use personal information such as demographics and socioeconomics, which are difficult to obtain online. As a solution, this study analyzes customers' interests and sentiments toward product features and uses them as customers' attributes. The stage consists of two steps: (i) identify product features mentioned in the online review data and (ii) extract customer attributes based on the feature-related cue phrases. Each step will be explained in the following subsections.

3.1.1 Identify Product Features in Online Data. In this step, the method for subfeature extraction suggested by Park and Kim [36] is used. It is considered necessary to summarize the method while the details are available in Ref. [36]. Figure 3 shows the flowchart of the method. First, the words in the review data are embedded into vectors by Word2Vec [37]. Next, phrases are extracted from the review data, and they are embedded into a vector space based on word vectors and product manuals. Finally, the phrase vectors are grouped into several clusters. The most frequent term in a cluster represents its subject, and feature-related clusters are manually selected based on the subject. The phrases in the selected ones represent subfeatures mentioned by customers.

3.1.2 Extract Customer Attributes. Next, cue phrases are extracted from the feature clusters. Then, the method analyzes customer attributes using these cue phrases. In this study, the customer attribute consists of feature+ and feature− to reflect both interests and sentiments regarding product features. The background is that customers who are satisfied with a feature F of product A and those complaining about F of the same product need to be distinguished. Therefore, the dimension of the attribute vector is twice the number of product features. There exist cases where both F^+ and F^- are 0 for a specific feature F: (i) a review does not contain cue phrases for F, e.g., *Absolutely love this phone* and (ii) a review mentions cue phrases for F with the neutral sentiment, e.g., *I got this on Prime Day for $100 off the price* for $F =$ price. The second case is different from no mention of the feature, but ignoring those reviews would be more appropriate for this study because the customer attribute is the combination of two properties—feature interest and sentiment. The reviews with neutral sentiments can be considered when the research focuses on feature interests only, as in Ref. [38]. Except for the above two cases, the method measures the sentiments of the sentences containing cue phrases for feature F and computes the average score. If the mean value is greater than 0, F^+ becomes 1. When the value is negative,

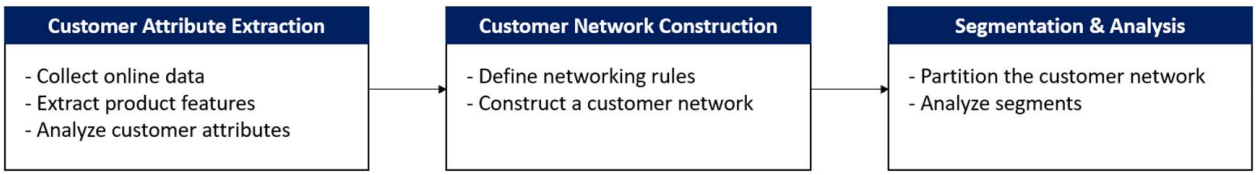

Fig. 2 Overview of the proposed methodology

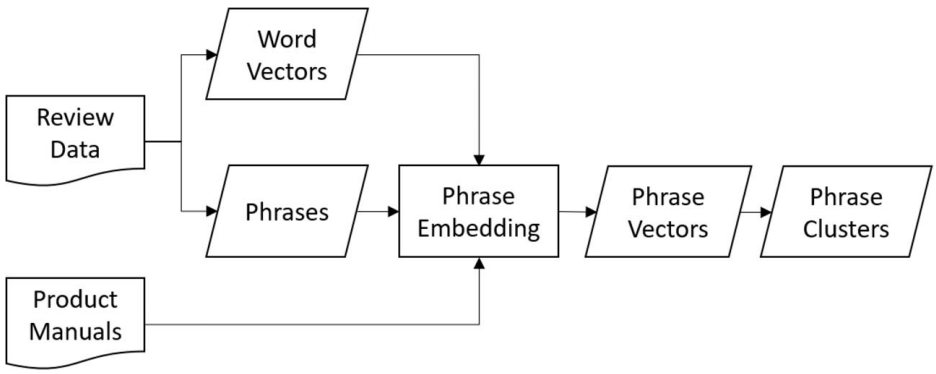

Fig. 3 Flowchart for subfeature extraction

Table 1 Extracting customer attributes

	Review 1				Great phone for the price. This phone is easy to use and feel like an expensive smart phone despite the cheap track phone price. The camera is nice and the interaction between command is […]					
	Review 2				Even with brightness at 100% the screen doesn't look the way it is supposed to					
	S+	S−	A+	A−	…	C+	C−	…	P+	P−
1	0	0	0	0	…	1	0	…	1	0
2	0	1	0	0	…	0	0	…	0	0

F^- becomes 1. Table 1 shows an example. Reviewer 1 expresses positive opinion about "price" which is a cue phrase for the price feature, so the attribute value for the price+ (P+) is 1. In the same way, the customer's attribute value for the camera+ (C+) is 1. Since no cue phrases for other product categories were found, the rest of the attributes are 0. In the second review, the customer complains about the brightness of "screen," a cue phrase for the screen feature. Therefore, Reviewer 2's attribute value for screen − (S−) is 1 and the others remain 0.

3.2 Customer Network Construction. In this stage, a customer network is constructed based on the customer attributes. Each customer becomes each node, and they are connected by predefined networking rules.

3.2.1 Define Networking Rules. There exist various approaches for connecting nodes in a network, such as mathematical rules [17] and qualitative relations between nodes [25]. The goal of this study is to group customers with similar attributes. Therefore, the proposed method uses the similarity between nodes for network construction. One of the popular methods for measuring similarity is cosine similarity. Equation (4) shows the definition of cosine similarity where $\vec{A}_i$ and $\vec{A}_j$ represent the attribute vector of customer i and j. It measures the cosine of the angle between two vectors projected in a multidimensional space

$$\mathrm{Sim}_C(i,j) = \frac{\vec{A}_i \cdot \vec{A}_j}{|\vec{A}_i| \times |\vec{A}_j|} \tag{4}$$

Although cosine similarity is applied in many disciplines [39–41], there are limitations when it comes to customer networks. One limitation is the lack of social meaning. Let us assume that the nodes are connected when $\mathrm{Sim}_C \geq 0.5$. This value of 0.5 can be numerically explainable. But it is not interpretable in terms of the social relationship among connected customers. In this study, the purpose of similarity measurement is to construct a customer network, a type of social network. Therefore, the measurement with social meanings is desirable. Another limitation is the quality of network clustering. In a small dimensional space (R^6), the network constructed by cosine similarity may not satisfy the clustering quality requirement [38].

As an alternative, this research proposes a new concept of "topic similarity," which measures the commonality of interests between customers. It is a relative concept, i.e., two customers may have different topic similarities. In Eq. (5), $\mathrm{Sim}_T^i(i,j)$ represents the topic similarity between customers i and j from customer i's perspective. It measures the ratio of the number of topics common in two customers to the number of topics mentioned by customer i. The second line of Eq. (5) shows the mathematical expression where a_k^i has a binary value and denotes the attribute value of customer i for topic k

$$\mathrm{Sim}_T^i(i,j) = \frac{\#\text{Topics common in customer } i,j}{\#\text{Topics mentioned by customer } i}$$
$$\mathrm{Sim}_T^i = \frac{\sum_{k=1}^{n} a_k^i a_k^j}{\sum_{k=1}^{n} a_k^i} \quad \text{and} \quad \mathrm{Sim}_T^j = \frac{\sum_{k=1}^{n} a_k^i a_k^j}{\sum_{k=1}^{n} a_k^j} \tag{5}$$

Equation (6) defines the rule for connecting two nodes in the network. If the similarity score is greater than or equal to a threshold value α for both customers, then two nodes corresponding to these customers are connected by an edge. Otherwise, the nodes cannot be connected

$$\mathrm{E}_{ij} = \begin{cases} 1 & \text{if } \mathrm{Sim}^i \geq \alpha \text{ and } \mathrm{Sim}^j \geq \alpha \\ 0 & \text{else} \end{cases} \tag{6}$$

Unlike cosine similarity, topic similarity gives the graph social meaning. For example, when we construct a customer network with the threshold $\alpha = 0.5$, it means that people with more than 50% similarity of interests are connected. The meaning of the "topic" may vary depending on the research domain. This article discusses customer segmentation for data-driven design based on user-generated data. Therefore, the topic means product features in this study.

3.2.2 Construct Customer Networks. Once the networking rules are determined, the method constructs customer networks by applying them to the extracted customer attributes. Let us assume that customers talk about six features (f1–f6), and the threshold value is 0.5. Figure 4 illustrates the process of network construction. First, the method calculates the topic similarity between customers 1 and 2. The number of common topics is 2 (f1, f6), so $\mathrm{Sim}_T^1(1,2) = \frac{2}{3} = 0.67$ and $\mathrm{Sim}_T^2(1,2) = \frac{2}{4} = 0.50$. Since both have similarity scores ≥ 0.5, two customers are connected. Next, it computes the similarity between customers 2 and 3. They have one common topic (f6), so $\mathrm{Sim}_T^2(2,3) = \frac{1}{4} = 0.25$ and $\mathrm{Sim}_T^3(2,3) = \frac{1}{2} = 0.50$. Two nodes cannot be connected because customer 2 does not satisfy the threshold. In the same way, customers 1 and 3 have similarity scores ≥ 0.5, and they are connected.

Both topic similarity and cosine similarity are used in this study to compare the results of modularity clustering. Since the network connects customers, it is a type of social networks. Therefore, the network should be a connected graph as real-world social networks are Ref. [42]. According to the rule in Eq. (6), a higher threshold means higher common interest between two customers. Therefore, the optimal threshold value is the highest α that constructs a connected graph. In Sec. 4, the optimal α will be determined by empirical analysis.

3.3 Segmentation and Analysis. In the final stage, the customer network is partitioned by modularity clustering. The resultant graph is analyzed to provide information about segments.

3.3.1 Partition Customer Networks. The modularity clustering stops iteration when the network has the maximum modularity score (Q). Therefore, the number of clusters is automatically determined. In the resultant network, each cluster represents a segment. Figure 5 shows an example of the modularity clustering result. The network is constructed based on the topic similarity among 200 customers. Modularity clustering divides the customers into five segments, each marked in a different color. The modularity score Q is 0.466. In the network clustering, the results with $Q \geq 0.3$ are

Customer	f1	f2	f3	f4	f5	f6	# Topics mentioned
1	1	0	0	1	0	1	3
2	1	1	0	0	1	1	4
3	0	0	0	1	0	1	2

Fig. 4 Network construction by topic similarity

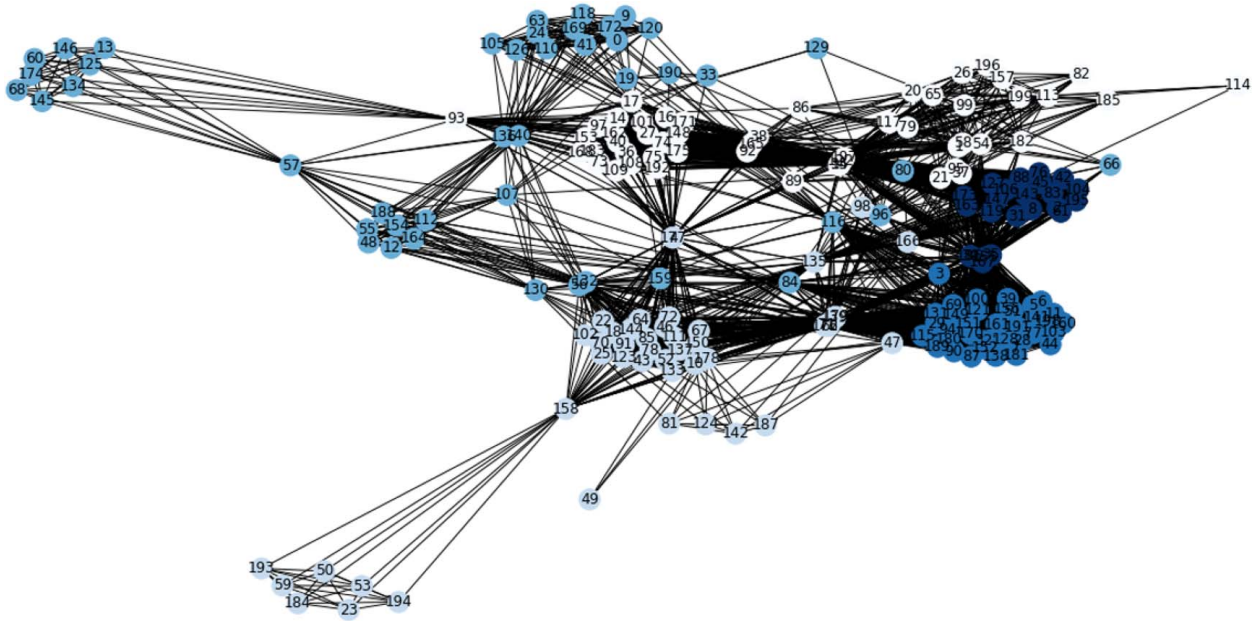

Fig. 5 Modularity clustering result ($N = 200$, $Q = 0.466$)

considered as good partitioning [43]. In Sec. 4, the goodness of segmentation results will be evaluated by Q.

3.3.2 Analyze Segments. The clustering result is analyzed to understand the characteristics of segments. In Sec. 3.1, a customer's attribute is defined based on the customer's interest and sentiment for product features. Accordingly, the segment property can be explained in members' interests and sentiments about product features. The concept of term frequency (TF) is adopted to measure this property. As TF represents the importance of a term within a set of documents [44], the attribute frequency can show the importance of an attribute within a group of customers. Therefore, a segment's property is measured by counting the frequency of attributes (feature ±) within the group.

Comparing the products purchased in each segment is another way to understand segment characters. Since a large number of products are used in data-driven customer analysis, the method adopts product clustering. First, product spec data are collected, and then the data are normalized by product features. For example, the largest screen size becomes 1, and the smallest screen size becomes 0. Next, the products are grouped by spectral clustering with K, the number of clusters automatically determined from HDBSCAN. Based on the resultant product clusters (PC), the method analyzes the sales record of PC in each customer segment and explains customers' purchasing behaviors.

4 Case Study

In this study, smartphone products are chosen for the case study because most people in US are familiar with product features with an 85% penetration rate [45].

4.1 Customer Attribute Extraction. In this stage, two types of datasets were collected: online customer reviews and product manual documents. A total of 25,340 reviews for 58 smartphone products were collected from Amazon.com. The reviews were written from May 2017 to July 2020. This study filtered reviews verified by Amazon for the authenticity of the data. The filtered data contain 109,688 sentences with 18,419 unique words. The average length of reviews is four sentences with 43 words. Regarding product manuals, documents for six smartphones were collected.

The collected data went through preprocessing. Special characters were removed, and end marks (?!.−) were replaced to a period. All letters were converted to lower cases. Non-English words were not removed since it excludes feature-related terms such as "GB" (unit for memory size) and "mAh" (unit for battery capacity). For data analysis, this study used a list of PYTHON libraries: (i) Spacy and Gensim for natural language processing and (ii) Hdbscan and Scikit-learn for data clustering. The words in review data were lemmatized and tokenized with Spacy. Then Gensim was used for word embedding. The parameters were set based on

Table 2 Cue phrases for product features

Screen	screen display, screen size, inch display, screen resolution, screen brightness, screen sensitivity, screen ratio, lcd screen, oled screen, screen clarity, huge screen, large screen, big screen, whole screen, screen edge, curved screen, etc.
AP[a]	fast processor, slow processor, snapdragon processor, exynos processor, process speed, processing speed
Memory	gb memory, storage capacity, internal memory, more memory, extra memory, expandable memory, gb ram, more storage, enough space, great storage, extra storage, internal storage, storage space, gb storage, more space
Camera	front camera, selfie camera, rear camera, main camera, mp camera, camera lens, camera quality, camera app, camera function, camera software, camera upgrade, well camera, camera shutter, camera sound, camera issue, camera noise, etc.
Battery	battery capacity, mAh battery, battery charge, battery life, battery percentage, battery saver, battery health, battery power, battery replacement, replaceable battery, removable battery, battery drain, low battery, large battery, battery level, etc.
Unlock	fingerprint reader, fingerprint sensor, fingerprint scanner, fingerprint reading, fingerprint recognition, finger print, finger scanner, finger reader, finger sensor, iris scanner, same finger, face recognition, facial recognition
Price	price range, price difference, price tag, decent price, affordable price, awesome price, perfect price, cheap price, excellent price, half price, retail price, amazing price, price drop, sale price, discount price, fair price, extra money, great value, etc.

[a]AP: application processor.

Ref. [9]. After training, Gensim returned a set of word vectors. For phrase extraction, the Noun_chunk and Textrank methods in Spacy were used. Phrases with a frequency less than 3 were deleted, and phrase embedding produced 1,969 phrases vectors. For phrase vector clustering, the hdbscan library and scikit-learn library in PYTHON were used. 108 clusters were obtained from each method, and resultant clusters were labeled based on the frequency analysis explained in Sec. 3.1.

Among 108 clusters from HDBSCAN clustering, 12 feature-related clusters were selected. Likewise, among 108 clusters from spectral clustering, eight feature-relevant clusters were chosen. Each cluster belongs to one of the seven feature categories shown in Table 2. The cue phrases for each feature were extracted from the corresponding clusters, and some technical terms were manually added based on the author's field experience in the smartphone industry. Using these cue phrases, this study analyzed customer attributes. At the end of this stage, each customer has an attribute vector of length 14. Since the method for feature identification is not limited to Ref. [36], the attribute vector can be shorter/longer than 14. For example, other approaches such as association rule mining [46] and LDA [47] may extract additional features such as color and weight and build longer attribute vectors. The resultant customer vectors of any length can be applied to the next stage.

4.2 Customer Network Construction. In online reviews, customers talk about various topics other than product features. They include customer services, returning products, and why they purchased products. Also, some customers wrote a simple review in one sentence. As a result, part of the reviews has 0 attributes since they did not mention product features, as shown below.

My daughter is an iPhone girl so we got this for her for graduation. It arrived on time and as described. It has worked well so far and my daughter loves it.

It worked perfectly.

In this study, the reviews with all 0 attributes were removed because they provide no implications for product features. Customer samples were taken from the remaining 8073 reviews. Then, a customer network was constructed based on the attributes and predefined networking rules. In this study, the networkx package in PYTHON was used for network construction. As mentioned in Sec. 3.2, the network should be a connected graph so that it resembles real-life social networks. According to Eq. (6), the graph structure is determined by the threshold value for connecting two nodes. To discover the optimal threshold value α for a connected graph, different α values were tested for randomly selected 200 reviews. Figure 6 shows the networks based on topic similarity with α values 0.4, 0.5, 0.6, and 0.7. In this empirical analysis, the highest α value that constructs a connected graph is 0.5. With $\alpha = 0.5$, the customer networks were constructed for different sample sizes. The minimum sample size was set to 980 by referring to the research about sample sizes in segment analysis [48]. Then, the customer network was constructed for the sample size 1K, 2K, to 8K.

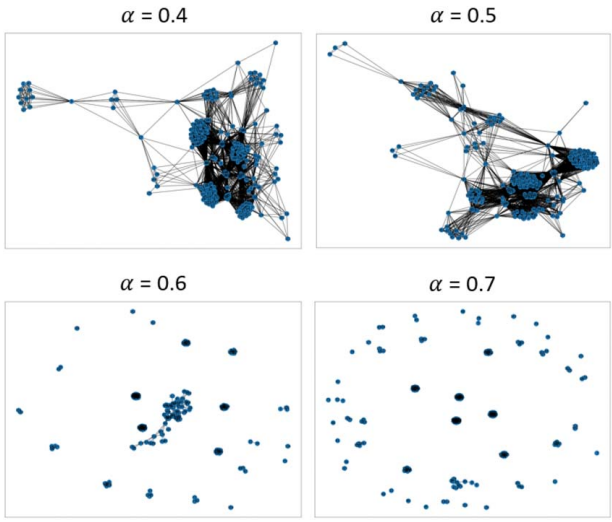

Fig. 6 Networks with different threshold values

4.3 Segmentation and Analysis. For customer segmentation, the constructed networks were partitioned by modularity clustering. In this research, the greedy_modularity_communities method from networkx in PYTHON was used. Figure 5 was created by this library. In the figure, it may seem questionable that node 93 belongs to the cluster on the center (Segment A), while it also has many connections

Table 3 Modularity scores

Sample size	Topic similarity	Cosine similarity
1000	0.443	0.371
2000	0.450	0.397
3000	0.448	0.376
4000	0.441	0.371
5000	0.446	0.366
6000	0.449	0.362
7000	0.444	0.367
8000	0.439	0.365

Note: $\alpha = 0.5$ for both methods.

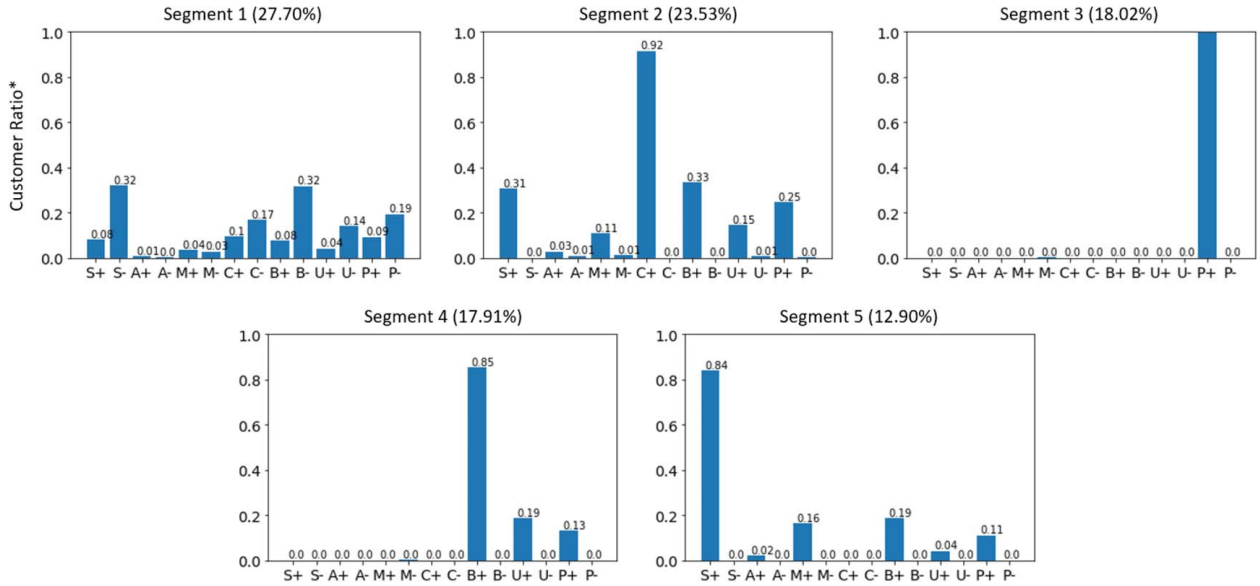

Fig. 7 Different characteristics of segments: Note * customer ratio: the ratio of customers with each attribute (feature + sentiment) within the segment, * values are rounded to two decimal places, M- in Segments 3 and 4 have values less than 0.005, so they are labeled as 0.0 in the graph.

with the top left one (Segment B). The analysis of raw data shows that Segment A and B have completely different attributes, and node 93 has similar but not the same attribute. According to the networking rule, nodes in Segment A and B cannot be connected, but node 93 is connected to both groups since it has the similarity score 0.5 for each group. In this case, the modularity score is the same no matter to which cluster node 93 belongs, so it is assigned to Segment A. Therefore, the result is acceptable in terms of clustering quality. This study conducted five trials for each sample size, and the average modularity scores are presented in Table 3. For the same threshold value ($\alpha = 0.5$), topic similarity provides better results than cosine similarity in clustering quality. The robustness of the results can be evaluated in two criteria: (i) the number of segments and (ii) the market share of each segment. From the sample size of 7000, the method provides consistent results in both criteria. In the following section, the segmentation result for the largest sample size (8000) is presented.

5 Results and Discussion

5.1 Segment Characteristics. This study started from the assumption that online customers have different preferences as offline customers do. The proposed methodology aims to divide online customers into segments that reflect heterogeneity among customers.

One way to examine this is to analyze customers' attributes in the resulting segments. Figure 7 shows the ratio of customers having each attribute within the group. For example, the value for S+ is calculated by (# customers expressing positive sentiments for the screen in the segment)/(# total customers in the segment). A customer mentioning multiple features is counted for each of those features. It shows that customers in different segments care about different product features and have nonidentical satisfaction levels for them. In segment 1, customers express negative sentiments for overall product features. Customers in Segment 2 mention positive opinions about most features with the highest interest in the camera feature. In Segment 3, customers focus on the price and have positive reviews. Customers in Segment 4 have positive feedback on the battery feature. In Segment 5, most customers show positive opinions about the screen feature.

The segments also can be characterized by the diversity of their interests. Figure 8 is the histogram of segmentation results. The *x*-axis indicates segment number, and the *y*-axis represents the number of feature mentioned in each review. It shows that segments have different diversity of interests. Specifically, customers in Segment 2 have the most diverse interests for product features, and some of them mention all features (seven feature categories). In Segment 1 and 4, most people talk about 1 or 2 features. People in Segment 3 and 5 focus on only one feature.

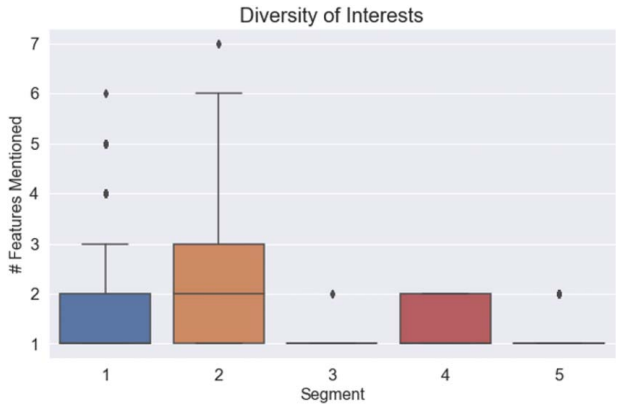

Fig. 8 Diversity of customer interests

5.2 Validation. The segmentation result can be validated by customer responses toward purchased products. For 58 products in the review data, their specifications were collected from websites focused on mobile devices (phonearena.com and gsmarena.com). These online sources provide detailed information about product features. Then, 58 products were clustered based on their specs. Table 4 shows the result with six PC. The spec range of products belonging to each PC is presented. Among features, the screen size and screen type have ordinal values. The higher value represents the higher quality. The unlock type feature has nominal values. In specific, a password is 0, fingerprint detection is 1, and face recognition is 2. The other features have continuous spec values. The last column shows the number of products.

Based on the clustering result, customer responses toward purchased products were analyzed. Figure 9 shows the ratio of PC in

Table 4 Product attributes in resultant product clusters

PC	S_{size}	S_{resol}	S_{type}	A_{speed}	A_n	M_{ram}	M_{rom}	C_{rear}	C_{front}	B_{cap}	U_{type}	*Price*	#
1	6.1–6.6	1,2	1,2,3	1.7–2.3	8	2–4	32–128	13–25	8–25	3500–5000	1	109–259	8
2	5.8–6.8	2,3	3	2.8	8	6–12	128–1024	12	8–16	3100–4500	1	349–849	11
3	4.7–6.5	1,2	2,3	2.3–2.7	4–6	2–4	32–64	12	7–12	1960–3969	1,2	208–949	9
4	6.2–6.7	2	2,3	2.0–2.8	8	4–6	64–128	12–108	16–32	3340–5260	1	206–449	12
5	5.5–6.3	2,3	3	2.0–2.8	8	4–6	64	12	8	2915–3500	1	269–499	9
6	4.7–6.4	1,2	1,2	1.4–2.4	8	1.5–4	16–64	5–16	2–8	1821–4080	0,1	119–299	9

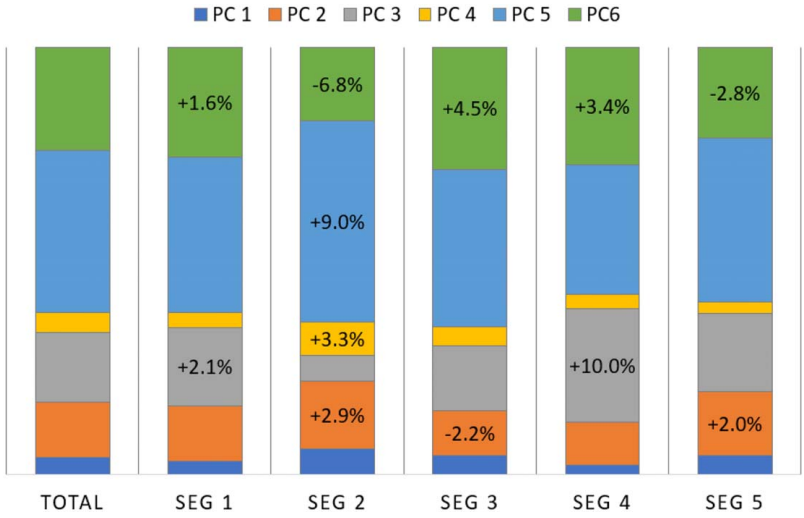

Fig. 9 The ratio of PC by segment

each segment. It is from the customer network with 8K samples and topic similarity. The first bar shows the ratio without segmentation which will be used as a baseline. For the others, the PC ratio varies according to segments. The changes can be explained in terms of segment characters (Figs. 7 and 8) and product specs (Table 4). SEG 1 is the only group where the majority of customers express negative opinions about product features. Specifically, the percentage of customers who mention negative sentiments toward at least one feature is: 99.95% (SEG 1), 3.13% (SEG 2), 0.26% (SEG 3), 0.21% (SEG 4), and 0.00% (SEG 5). Therefore, an increased PC ratio in SEG 1 means that customers express more complaints about that PC than other product groups. According to Fig. 7, SEG 1 cares about the screen, camera, battery, and price. PC 3 has relatively low specs for these features and includes the most expensive products, which leads to lower satisfaction. As a result, customers express more complaints about PC3, and it has a higher ratio in SEG 1 than the baseline. PC 6 has the lowest spec values in the overall product attributes, so it also has an increased ratio. The rest of the segments mention positive opinions about product features. Therefore, an increased PC ratio means that customer express more compliments about that PC with higher satisfaction. Figure 7 shows that customers in SEG 2 are highly interested in the camera. PC 5 contains Google pixel phones focused on camera features, so it gains many compliments from users. PC 4 has the highest specs for front and rear cameras and obtains more positive responses from customers. On the other hand, PC 6 with the lowest camera specs has less positive feedback resulting in a decreased PC ratio. In SEG 2, customers also care about the screen feature. Since they are interested in several features, as shown in Fig. 8, the customers may consider the camera and screen together. The screen subfeatures related to the camera are resolution and type. Therefore, PC 2 with the highest screen resolution and type have an increased ratio. SEG 3 focuses on the price. As a result, PC 6 with low-tier products has more positive reviews resulting in an increased PC ratio while PC 2 and 3 containing the highest prices have a decreased ratio. In SEG 4, customers have a high interest in the battery feature. Interestingly, PC 3 and 6 with the lowest battery capacity have an increased ratio, and other PC have a decreased ratio. This result implies that what customers care about is battery usage time rather than battery capacity. Unlike general spec values, the usage time is hard to collect because it is obtained from simulations such as call, video playing, and internet surfing [49]. Additional analysis on usage time will be considered in future works. In SEG 5, customers are interested in the screen feature. PC 2 has the highest specs in overall screen features, so it has an increased ratio resulting from more positive reviews. PC 6 contains the lowest specs in all screen subfeatures and has a decreased ratio in SEG 5.

The validity of the proposed methodology can also be shown by comparing the network clustering with previous methods. In the previous studies on customer segmentation, nonhierarchical methods such as K-means clustering are the most popular methods [3]. They produced proper results for the customer attributes based on continuous variables such as demographics and socioeconomic factors. However, for online customers, the attribute

Table 5 X-means clustering result

Cluster	Item	Count
1	[0 0 0 0 0 0 0 0 0 0 0 0 1 0]	1431
2	[0 0 1 0 0 0 0 0 0 0 0 0 1 0]	1
3	[0 0 0 0 0 1 0 0 0 0 0 0 1 0]	3
4	[0 0 0 0 0 0 1 0 0 0 0 0 1 0]	130
5	[0 0 1 0 0 0 1 0 0 0 0 0 1 0]	5
…	…	…
233	[0 1 0 0 0 0 1 0 1 0 0 0 0 0]	7

Table 6 Customer utility evaluation

	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6
Screen size (inch)	6.2	6.8	6.5	6.2	5.8	5.0
AP speed (GHz)	2.1	2.8	2.5	2.2	2.8	1.6
Memory rom (GB)	32	256	64	128	64	16
Camera rear (MP)	13	12	12	24	12	5
Battery capacity (mAh)	4000	4300	3179	3340	3000	2600
Unlock type	1	1	2	1	1	1
Price (USD)	130.00	679.99	624.95	229.50	499.99	119.99
Segment 1	0.606	0.604	0.482	0.536	0.259	0.179
Segment 2	0.527	0.537	0.434	0.682	0.299	0.117
Segment 3	0.903	0.005	0.022	0.421	0.079	0.994
Segment 4	0.710	0.883	0.375	0.450	0.243	0.087
Segment 5	0.554	0.679	0.775	0.500	0.311	0.112
Unsegmented	0.628	0.563	0.421	0.553	0.253	0.235

is their interest in product features. Since it is a vector consisting of binary values, the customer attribute space has a very sparse distribution. As a result, K-means clustering does not work properly for online customers. In this study, X-means clustering [50] was applied to the randomly selected 8000 sample customers. The sample contains 345 different attributes, and X-means resulted in 233 clusters, as shown in Table 5. The item column shows the customer attribute in each cluster, and the count column indicates the number of customers. Most clusters have a single attribute, i.e., all vectors within a cluster are the same, which is not a proper result for customer segmentation. The proposed methodology addressed this problem by connecting customers with similar tendencies and then partitioning the whole customer base.

5.3 Design Application. In this section, the practical application of the proposed methodology is presented with a simulation. A company usually develops several design concepts and evaluates them by surveys. The number of product alternatives in surveys ranges from 3 to 6 [51], so this study tested six design candidates, as shown in Table 6. Each candidate is randomly selected from different product clusters. Let us assume that these are design candidates for the company's new product. They can be evaluated by the random utility model in Eq. (7), where U_{ni} represents the utility of customer n obtained by purchasing product i

$$U_{ni} = V_{ni} + \varepsilon_{ni} = \sum_k \beta_{nk} x_{ik} + \varepsilon_{ni} \tag{7}$$

The deterministic part V_{ni} is a weighted sum of product features, and ε_{ni} is a random error term [52]. In Eq. (7), β_{nk} is the weight that customer n has for product feature k, and x_{ik} is the spec value for feature k of product i. Based on the principle of utility maximization [53], customers will choose the design option with the highest utility. Therefore, when the company plans a new product, its priority goes to the design candidate with the highest utility.

The utility value is dependent on the weight β_{nk}. There are different approaches for extracting feature weight from online data, such as choice models [19] and neural networks [14]. For the simplicity of simulations, this study adopts the method of Kim et al. [54]. They used the term frequency (TF) to extract the partial utility of product components from customer reviews. The approach is based on the assumption that more frequently mentioned terms have higher importance because people have more interest in them. This study modifies the TF analysis used in Sec. 5.1 to obtain the weight for product feature k. First, in each segment, the frequency of each feature is counted. Table 7 shows the resultant TF for Segment 2. Since some segments contain 0 value of TF, the offset of 1 is applied to the initial TF. Then, the TF ratio is used as the feature weight so that the total weights are summed up to 1. The weights for the other segments are obtained in the same way. There are different ways to analyze the feature importance, and they also can be used for the deterministic part V_{ni}.

Table 7 Weight for product features (Segment 2)

	TF	Offset	Weight
Screen	580	581	0.145
AP	72	73	0.018
Memory	226	227	0.057
Camera	1731	1732	0.432
Battery	633	634	0.158
Unlock	290	291	0.073
Price	469	470	0.117
Total		4008	1.000

Based on these weights, the customer utility for design candidates is evaluated. It is assumed that customers prefer lower prices, so the price attribute is $x_{ki} = 1/\text{price}$. In Table 6, the second table shows the computed utility of each design for five segments and the entire customers. The highest utility within a group is highlighted in blue. The result shows that different segment has a different preference for the suggested design options. To be specific, Segment 1 prefers Design 1, and Segment 2 prefers Design 4. For Segment 3, the priority goes to Design 6. Segment 4 has the highest utility for Design 2, and Segment 5 prefers Design 3. All segments have a different first choice. Without segmentation, the company would choose Design 1, but this only matches Segment 1. This example shows that the proposed methodology can help companies design target-customer-oriented products by reflecting their distinguished preferences for product features.

6 Conclusion and Future Works

This study focused on the gap between field and academic research in customer analysis. With increasing online platforms and the development of data analysis techniques, online data have emerged as an efficient resource for customer analysis. The online data have strength in that it is time and cost-efficient compared to surveys. But most studies about online customer analysis neglected segmentation. They treated online customers as if all customers have similar characters and preferences. However, in the field, customers are considered to have different tendencies, so customer segmentation is the key strategy in product design and marketing today [3].

This paper proposed a new methodology for online customer segmentation to close this gap. First, the method extracted product features from online review data and identified related cue phrases. Then, it analyzed each customer's interest in product features using these cue phrases. The result became the customer's attributes. Second, the method measured the similarity between customers based on the extracted customer attributes and two indices:

cosine similarity and topic similarity. By connecting customers with the similarity score above a threshold, the method constructed a customer network. Finally, this network was partitioned by modularity clustering. The resultant segments were analyzed in three criteria: (i) interests for product features; (ii) diversity of interests; and (iii) responses toward purchased products. The result showed that the proposed methodology properly reflects the heterogeneity of the online customer base in the segmentation result. Also, an example of a design application was presented. It showed how the suggested method can help companies design target-customer-oriented products.

In future works, segments will be further analyzed to get implications for product design. For example, users may have different feature expectations by tier. They would expect different spec levels for flagship smartphones and affordable smartphones. The effect of price on each segment will be further investigated in future works. Other design implications such as feature importance will be studied as well. As mentioned in Sec. 2, there have been many studies on online customer analysis. Some of the previous works can be adopted for segment analysis, and the result will provide customized design implications. The effect of customer segmentation can be demonstrated by comparing the obtained design implications with and without segmentation.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

Appendix

Table 8 shows the description of product features in Table 4.

Table 8 Description of features in Table 4

Notation	Feature	Unit
S_{size}	Screen size	Inch
S_{resol}	Screen Resolution	HD = 1, FHD = 2, QHD = 3
S_{type}	Screen Type	TFT = 1, IPS = 2, OLED = 3
A_{speed}	Application Processor Speed	GHz
A_n	Number of Application Processor Cores	
M_{ram}	Memory RAM	GB
M_{rom}	Memory ROM	GB
C_{rear}	Rear Camera	MP
C_{front}	Front Camera	MP
B_{cap}	Battery Capacity	mAh
U_{type}	Unlock type	Passcode = 0, Fingerprint = 1, Face ID = 2
Price	Price	USD

References

[1] Kotler, P., and Keller, K. L., 2016, *A Framework for Marketing Management*, Pearson, Harlow.
[2] Camilleri, M. A., 2018, *Travel Marketing, Tourism Economics and the Airline Product: An Introduction to Theory and Practice*, Springer, Cham.
[3] Tuma, M. N., Decker, R., and Scholz, S. W., 2011, "A Survey of the Challenges and Pifalls of Cluster Analysis Application in Market Segmentation," Int. J. Market Res., **53**(3), pp. 391–414.
[4] Fink, A., 2002, *How to Ask Survey Questions*, SAGE, Thousand Oaks, CA.
[5] Kim, H. H. M., Liu, Y., Wang, C. C., and Wang, Y., 2017, "Data-Driven Design (d3)," ASME J. Mech. Des., **139**(11), p. 110301.
[6] Joung, J., and Kim, H. M., 2021, "Automated Keyword Filtering in Latent Dirichlet Allocation for Identifying Product Attributes From Online Reviews," ASME J. Mech. Des., **143**(8), p. 084501.
[7] Tuarob, S., and Tucker, C., 2015, "Quantifying Product Favorability and Extracting Notable Product Features Using Large Scale Social Media Data," ASME J. Comput. Inf. Sci. Eng., **15**(3), p. 031003.
[8] Yang, B., Liu, Y., Liang, Y., and Tang, M., 2019, "Exploiting User Experience From Online Customer Reviews for Product Design," Int. J. Inform. Manage., **46**, pp. 173–186.
[9] 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.
[10] Smith, W. R., 1956, "Product Differentiation and Market Segmentation As Alternative Marketing Strategies," J. Marketing, **21**(1), pp. 3–8.
[11] Liang, T.-P., and Liu, Y.-H., 2018, "Research Landscape of Business Intelligence and Big Data Analytics: A Bibliometrics Study," Expert Syst. Appl., **111**, pp. 2–10.
[12] Zhou, F., Ayoub, J., Xu, Q., and Yang, X. J., 2019, "A Machine Learning Approach to Customer Needs Analysis for Product Ecosystems," ASME J. Mech. Des., **142**(1), p. 011101.
[13] Zhang, D., Xu, H., Su, Z., and Xu, Y., 2015, "Chinese Comments Sentiment Classification Based on Word2vec and Svmperf," Expert Syst. Appl., **42**(4), pp. 1857–1863.
[14] Joung, J., and Kim, H. M., 2021, "Approach for Importance–Performance Analysis of Product Attributes From Online Reviews," ASME J. Mech. Des., **143**(8), p. 081705.
[15] Tuarob, S., and Tucker, C. S., 2015, "Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks," ASME J. Mech. Des., **137**(7), p. 071402.
[16] 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.
[17] Wang, M., and Chen, W., 2015, "A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design," ASME J. Mech. Des., **137**(7), p. 071410.
[18] Park, Y., and Lee, S., 2011, "How to Design and Utilize Online Customer Center to Support New Product Concept Generation," Expert Syst. Appl., **38**(8), pp. 10638–10647.
[19] Suryadi, D., and Kim, H. M., 2019, "A Data-Driven Methodology to Construct Customer Choice Sets Using Online Data and Customer Reviews," ASME J. Mech. Des., **141**(11), p. 111103.
[20] Bondy, A. J., 1976, *Graph Theory With Applications*, Macmillan, London.
[21] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G., 2009, "Network Analysis in the Social Sciences," Science, **323**(5916), pp. 892–895.
[22] Derrible, S., and Kennedy, C., 2009, "Network Analysis of World Subway Systems Using Updated Graph Theory," Transport Res. Record: J. Transport Res. Board, **2112**(1), pp. 17–25.
[23] Xing, W., and Ghorbani, A., 2004, "Weighted Pagerank Algorithm," Proceedings. Second Annual Conference on Communication Networks and Services Research, Fredericton, NB, Canada, May 21, IEEE, pp. 305–314.
[24] Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M., 2012, "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," Marketing Sci., **31**(3), pp. 521–543.
[25] Sosa, M. E., Eppinger, S. D., and Rowles, C. M., 2007, "A Network Approach to Define Modularity of Components in Complex Products," ASME J. Mech. Des., **129**(11), pp. 1118–1129.
[26] Jamali, M., and Abolhassani, H., 2006, "Different Aspects of Social Network Analysis," 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06), Hong Kong, China, Dec. 18–22, pp. 66–72.
[27] Das, K., Samanta, S., and Pal, M., 2018, "Study on Centrality Measures in Social Networks: A Survey," Soc. Netw. Anal. Min., **8**(1), pp. 1–11.
[28] Hao, F., Min, G., Pei, Z., Park, D., and Yang, L. T., 2017, "*k*-clique Community Detection in Social Networks Based on Formal Concept Analysis," IEEE Syst. J., **11**(1), pp. 250–259.
[29] Emmons, S., Kobourov, S., Gallant, M., and Börner, K., 2016, "Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale," PLoS One, **11**(7), p. e0159161.
[30] Von Luxburg, U., 2007, "A Tutorial on Spectral Clustering," Stat. Comput., **17**(4), pp. 395–416.
[31] Girvan, M., and Newman, M. E., 2002, "Community Structure in Social and Biological Networks," Proc. Natl. Acad. Sci. USA, **99**(12), pp. 7821–7826.
[32] Blondel, V. D., Guillaume, J. -L., Lambiotte, R., and Lefebvre, E., 2008, "Fast Unfolding of Communities in Large Networks," J. Stat. Mech.: Theory Exp., **2008**(10), p. P10008.
[33] Dinh, T. N., Li, X., and Thai, M. T., 2015, "Network Clustering Via Maximizing Modularity: Approximation Algorithms and Theoretical Limits," 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, Nov. 14–17, IEEE, pp. 101–110.
[34] Sánchez, D. L., Revuelta, J., Prieta, F. D. L., Gil-González, A. B., and Dang, C., 2016, "Twitter User Clustering Based on Their Preferences and the Louvain Algorithm," International Conference on Practical Applications of Agents and Multi-agent Systems, Sevilla, Spain, June 1–3, Springer, pp. 349–356.
[35] Rahiminejad, S., Maurya, M. R., and Subramaniam, S., 2019, "Topological and Functional Comparison of Community Detection Algorithms in Biological Networks," BMC Bioinform., **20**(1), pp. 1–25.

[36] Park, S., and Kim, H. M., 2022, "Phrase Embedding and Clustering for Sub-feature Extraction From Online Data," ASME J. Mech. Des., **144**(5), p. 054501.
[37] Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013, "Efficient Estimation of Word Representations in Vector Space". e-print arXiv:1301.3781v.
[38] Park, S., and Kim, H. M., 2021, "Data-Driven Customer Segmentation Based on Online Review Analysis and Customer Network Construction," International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 85383, American Society of Mechanical Engineers, Paper No. V03AT03A015.
[39] Ye, J., 2011, "Cosine Similarity Measures for Intuitionistic Fuzzy Sets and Their Applications," Math. Comput. Modell., **53**(1–2), pp. 91–97.
[40] Nguyen, H. V., and Bai, L., 2011, "Cosine Similarity Metric Learning for Face Verification," Asian Conference on Computer Vision, Queenstown, New Zealand, Nov. 8–12, pp. 709–720.
[41] Li, J., Wu, Z., Zhu, B., and Xu, K., 2018, "Making Sense of Organization Dynamics Using Text Analysis," Expert Syst. Appl., **111**, pp. 107–119.
[42] Newman, M. E. J., Watts, D. J., and Strogatz, S. H., 2002, "Random Graph Models of Social Networks," Proceedings of the National Academy of Sciences, Vol. 99, pp. 2566–2572.
[43] Fortunato, S., and Barthélemy, M., 2007, "Resolution Limit in Community Detection," Proc. Natl. Acad. Sci. USA, **104**(1), pp. 36–41.
[44] Manning, C. D., Raghavan, P., and Schütze, H., 2009, *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK.
[45] O'Dea, S., 2021, "Smartphones in the U.S. – Statistics & Facts," https://www.statista.com/topics/2711/us-smartphone-market/
[46] Hu, M., and Liu, B., 2004, "Mining Opinion Features in Customer Reviews," 19th National Conference on Artificial Intelligence, San Jose, CA, July 25–29, pp. 755–760.
[47] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, "Latent Dirichlet Allocation," J. Mach. Learn. Res., **3**, pp. 993–1022.
[48] Dolnicar, S., Grün, B., Leisch, F., and Schmidt, K., 2013, "Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism," J. Travel Res., **53**(3), pp. 296–306.
[49] Michaels, P., 2021, "Best Phone Battery Life in 2022: The Longest Lasting Smartphones," https://www.tomsguide.com/us/smartphones-best-battery-life,review-2857.html
[50] Pelleg, D., and Moore, A. W., 2000, "X-means: Extending K-means With Efficient Estimation of the Number of Clusters," ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, June 29–July 2, pp. 723–734.
[51] Shocker, A. D., Ben-Akiva, M., Boccara, B., and Nedungadi, P., 1991, "Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models, and Suggestions," Marketing Lett., **2**(3), pp. 181–197.
[52] Chen, W., Hoyle, C., and Wassenaar, H. J., 2013, *Decision-Based Design*, Springer, London.
[53] Ben-Akiva, M., and Lermna, S. R., 1985, *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, Cambridge.
[54] Kim, J., Park, S., and Kim, H. M., 2022, "Optimal Modular Remanufactured Product Configuration and Harvesting Planning for End-of-Life Products," ASME J. Mech. Des., **144**(4), p. 042001.