ASME Journal of Mechanical Design
Online journal at:
https://asmedigitalcollection.asme.org/mechanicaldesign

Customer Segmentation and Need Analysis Based on Sentiment Network of Online Reviewers and Graph Embedding

Mengyuan Shen[1]
University of Michigan—Shanghai Jiao Tong
University Joint Institute,
Shanghai Jiao Tong University,
800 Dongchuan Road, Minhang District,
Shanghai 200240, China
e-mail: shenmengyuan@sjtu.edu.cn

Bohan Feng[1]
University of Michigan—Shanghai Jiao Tong
University Joint Institute,
Shanghai Jiao Tong University,
800 Dongchuan Road, Minhang District,
Shanghai 200240, China
e-mail: bohan.feng@sjtu.edu.cn

Aoxiang Cheng
University of Michigan—Shanghai Jiao Tong
University Joint Institute,
Shanghai Jiao Tong University,
800 Dongchuan Road, Minhang District,
Shanghai 200240, China
e-mail: aoxiang.cheng@sjtu.edu.cn

Youyi Bi[2]
University of Michigan—Shanghai Jiao Tong
University Joint Institute,
Shanghai Jiao Tong University,
800 Dongchuan Road, Minhang District,
Shanghai 200240, China
e-mail: youyi.bi@sjtu.edu.cn

Customer segmentation divides customers into groups with different characteristics and supports the design of customized products and tailored marketing strategies. Recent studies explore using online reviews as the data source and social network analysis as the fundamental technique for customer segmentation. These studies usually utilize the frequency of mentioned product attributes and/or customers' sentiments from online reviews in the segmentation process. However, few of them investigate the influence of different types of information (e.g., with or without sentiment, order information) on the segmentation performance. In addition, previous studies seldom consider and tackle the challenge of clustering high-dimensional data when online reviews contain customers' rich opinions towards multi-faceted attributes of a product. To fill these gaps, we propose a comprehensive framework for customer segmentation and need analysis based on sentiment network of online reviewers and graph embedding. The frequently mentioned product attributes and customers' sentiments are first extracted from online reviews. Then, a customer can be represented as a vector consisting of his/her sentiment polarities on each product attribute as well as rating and order information. After that, a social network of customers is established by examining the similarity of customer vectors. The network nodes are embedded into low-dimensional vectors, which can be further clustered into different groups, i.e., customer segments, and their respective needs can be analyzed by methods such as Importance–Performance Analysis. Our framework enables the construction and performance comparison of various types of networks, node compositions, and embedding methods. A case study employing the online reviews of a passenger vehicle in China's market is used to demonstrate the validity of the proposed framework. The results indicate that the customer segmentation generated by the sentiment network of online reviewers with Graph Autoencoder (GAE) embeddings performs better than other alternative models that do not utilize vector embeddings, fail to consider the sentiment information, or leverage bipartite network structures. Our framework provides more nuanced insights for designers to improve customers' satisfaction and increase the market competitiveness of their products.

[DOI: 10.1115/1.4067226]

Keywords: customer segmentation, online review, social network, graph embedding, sentiment analysis, data-driven design, design for market systems

1 Introduction

Customer segmentation divides customers into groups with different characteristics and needs according to their behaviors, individual attributes, and preferences [1]. It has been widely used by designers and marketers to gain insights from the features of each customer segment and improve the design of products and marketing strategies [2,3]. For instance, as a popular strategic approach in modern marketing, the STP (Segmentation, Targeting, Positioning) marketing model [4] sets segmentation as its first step. The main goal is to create various customer segments based on customers' features. Usually, the demographic, psychographic, and behavioral features of customers are the three most significant dimensions for customer segmentation [5]. After segmentation, companies can determine the target segments and position the product or service within the market accordingly.

Traditional customer segmentation research primarily utilizes surveys or sale records as the data source. However, the collection of such data usually requires significant time and financial resources. Recent studies [6–8] explore using online reviews as they can be collected faster and less expensively. The ease of accessing online reviews makes it possible to collect large-scale

[1] Mengyuan Shen and Bohan Feng contributed equally to this paper.

[2] Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received July 2, 2024; final manuscript received October 27, 2024; published online December 13, 2024. Assoc. Editor: Cyril Picard.

data in real time, which can better reflect market dynamics longitudinally. In addition, researchers also explore constructing social networks of customers by extracting critical information from online reviews for customer segmentation [9]. Customer social networks can be built based on the similarity of online reviewers, and then the clustering of customers is performed on the network. The development of network-based methods can better capture the interrelationships among customers for customer segmentation.

However, several critical gaps still exist in the research on network-based customer segmentation. First, online reviews usually contain the occurrences of mentioned product attributes, customers' sentiment towards products as well as ratings and order information. Few existing studies investigate the influence of using different types of information (e.g., with or without sentiment and order information) from online reviews on segmentation performance. Second, the dimension of the node attribute vector in a customer social network built from online reviews can be quite high as online reviews may include rich information regarding customer's opinions on multiple attributes of a product, and clustering high-dimensional data is challenging due to the "curse of dimensionality" [10]. The similarity measures become less meaningful as dimensions of clustered points increase, making regular clustering algorithms ineffective [11]. Third, the performance of different network architectures for customer segmentation is under-researched. It is still unclear how network types (e.g., homogenous or bipartite, weighted or unweighted) influence the effect of network-based customer segmentation. Last, a comprehensive framework of customer segmentation and need analysis based on customers' social networks built from online reviews is lacked.

To fill these gaps, we propose a framework for customer segmentation and need analysis based on sentiment network of online reviewers and graph embedding. In this framework, homogeneous and bipartite networks are built with different customer features extracted from online reviews. Then, these networks are embedded, and the obtained low-dimensional vectors of nodes are clustered into customer segments. This treatment can capture the complex and latent interrelationships and nuances among customers in the segmentation process. The final network model and obtained segments are selected based on the clustering performance. After the customer segmentation, product attribute analysis methods such as Importance–Performance Analysis (IPA) can be applied to each segment to analyze customer needs and provide corresponding suggestions for design improvement. To demonstrate the proposed framework, the online reviews of a popular passenger vehicle from China's automotive market are utilized in a case study. The results indicate that the customer segmentation generated by the homogenous sentiment network of online reviewers with Graph Autoencoder (GAE) embeddings performs better than other alternative models that do not utilize vector embeddings, fail to consider the sentiment information, or leverage bipartite network structures. Our framework can support designers to improve customer satisfaction and market competitiveness of their products by extracting nuanced insights from online reviews.

The rest of this paper is organized as follows. Section 2 reviews recent progress in customer segmentation, especially the application of network-based methods. Section 3 explains the proposed framework and related key methods, including data collection and preprocessing, product attributes identification, construction and embedding of customer social network, customer segmentation, and need analysis. Section 4 presents a case study and discusses the customer segmentation results. Section 5 summarizes our research contribution and findings, and provides recommendations for future work.

2 Related Work

Customer segmentation can reflect the variances of customer needs and is critical to the success of product design in dynamic and diversified markets. Researchers from engineering design and marketing have achieved rich results in this area. For example, Smith et al. [12] chose an interactive version of the genetic algorithm to simultaneously discover optimal multi-attributed products for different customer segments. Ertian et al. [13] proposed a method for customer demand segmentation based on fuzzy clustering and trigonometric functions so that manufacturers can meet the demands of different customers with fewer product designs. Hu et al. [14] developed a framework where explainable artificial intelligence is adopted to enhance designers' trust in AI predictions for customer segmentation in product development.

Traditional customer segmentation studies mainly reply on data sources such as surveys and sale records. For example, Wu and Chou [15] combined a soft clustering method and the Latent Dirichlet Allocation (LDA) model to divide online customers into different segments based on online questionnaires. Then, the shopping behaviors, product satisfaction levels, and demographic characteristics of customers are summarized for each segment to guide companies in improving their marketing strategies. Peker et al. [16] applied the *k*-means clustering method on sale records from a chain grocery store and utilized the LRFMP model (Length, Recency, Frequency, Monetary, and Periodicity) to categorize customers into groups with different profiles and brand loyalty levels. This categorization can be used to guide the development of the grocery industry based on different customer groups' behaviors and preferences.

In recent years, online reviews have emerged as a significant type of data source for customer segmentation due to their lower accessing cost and better timeliness. For example, Wang et al. [17] developed a systematic methodology for eliciting product attributes from User-Generated Content (e.g., online customer reviews, blogs, and social networking interactions), constructing customer preference models and using these models in design selection. Jiang et al. [18] proposed a customer segment analysis approach based on online customer reviews of durable products. Their approach considers reviewers' mention of product features, and the probability-based LCA (latent class analysis) method is adopted upon the characteristics of online reviews, to effectively cluster reviewers into specified segmentations. The segment analysis result can provide support for new product design and development, repositioning of existing products, marketing strategy development, and product differentiation. Joung and Kim [19] proposed an interpretable machine learning-based approach for customer segmentation for new product development based on the importance of product features from online product reviews. Their approach can identify a group of customers with unsatisfied needs and support the development of new product concepts.

Furthermore, researchers explore constructing a social network of customers from online reviews to better support customer segmentation. As a classical data structure consisting of a set of nodes and relationships (edges) connecting nodes [20], networks can represent complex relations in many real-world scenarios and have been utilized in customer preference modeling [21,22] and customer segmentation analysis [9]. For example, Wang [23] proposed a network analysis approach for market segmentation of online reviews, which considers the networked nature of interactive relationships among reviewers and brands across online review websites using core-periphery structure and centrality measures. Helal et al. [24] proposed a novel social network mining approach which detects communities based on the most influential users of a specific social network with a real-world social network dataset where online users rate movies. They used a direct data mining approach based on frequent pattern discovery for discovering leaders. Those leaders are then used as core members to expand the communities around them. Park and Kim [9] constructed a one-dimensional customer social network based on the similarity of the attributes mentioned by customers from their online reviews. Then, the clustering of customers is performed on the network nodes. Their study contributes to the development of network-based methods to capture interrelationships among customers for customer segmentation.

Existing studies usually utilize the frequency of mentioned product attributes and/or customers' sentiments from online reviews to represent customers in the segmentation process. For instance, a typical representation of customers is using vectors consisting of product features with sentiment polarity (positive/negative) [8,9]. However, few existing studies investigate the influence of using different types of information from online reviews (e.g., with or without sentiment, order, and rating information) on the segmentation performance. The lack of these insights may lead to biased customer segmentation results. In addition, as more information is included in the node attribute vectors of customers' social network, clustering these high-dimensional nodes becomes more and more challenging.

An emerging kind of method for reducing the dimensionality of graph data while keeping their key features in complex interrelationships is graph embedding. After the network construction, graph embedding can be applied to obtain low-dimensional vectors of network nodes. Common graph embedding methods contain Graph Neural Network (GNN) [25], Graph Convolutional Network (GCN) [26], Graph Attention Network (GAT) [27], and Graph Sample and Aggregate (graphSAGE) [28]. Graph embedding methods have also been developed for bipartite and heterogeneous networks, such as Heterogeneous Network Embedding (HNE) [29], Metapath2vec++ [30], and Bipartite Network Embedding (BiNE) [31]. These graph embedding models aggregate the information from neighboring nodes to update node features. Although graph embedding has been applied in sentiment analysis of online reviews [32,33] and the design of e-commerce recommendation systems [34], few studies explored the integration of network-based methods and graph embedding in customer segmentation for product design.

To fill these gaps, in this study, we expect to construct sentiment network of online reviewers and generate low-dimensional representations of customers (i.e., embedding vectors) for customer segmentation. In addition, the influence of different network architectures and node compositions on the performance of customer segmentation will be examined and compared.

3 Methodology

3.1 Overall Structure. Figure 1 shows the overall flow of the proposed framework consisting of three major stages: (1) Data Processing and Information Extraction, (2) Network Construction and Embedding, and (3) Customer Segmentation and Need Analysis. In Stage 1, the collected online reviews are preprocessed, and those frequently mentioned product attributes are extracted from online reviews through keyword extraction after sentence segmentation and part-of-speech tagging. An attribute-keyword dictionary is established to support this process, which can be created by mapping the extracted keywords with corresponding components or functions of the product (e.g., mapping the keyword "power" with the engine system of a vehicle). After that, customers' sentiment intensities for product attributes are calculated and classified into three polarities (i.e., positive, neutral, and negative). Then, a customer can be represented by a sentiment vector consisting of his/her sentiment polarities on each product attribute. Online reviews usually also include customer ratings on products and their order information such as purchase price, location, and purpose. This information can also be extracted and converted into vectors. Customers' sentiment vectors, product rating vectors, and order information vectors are combined as customer vectors. In addition to customer vectors, product attribute vectors are also generated using the calculated occurrence frequencies of the corresponding keywords of the attributes in collected online reviews. Both customer vectors and product attribute vectors will be used for network construction in Stage 2.

In Stage 2, customer networks are established with different choices of network structures, edge weights, and node attributes. Commonly adopted network structures include homogeneous networks and bipartite networks. In homogeneous networks, nodes represent customers posting online reviews and links represent their connections. The formation of a link depends on whether the similarity of two customer vectors is above a threshold. In bipartite networks, nodes represent customers and product attributes, and the links represent customers' commenting frequency on the attributes. These links can also be weighted with customer sentiment polarities. The constructed networks are then put into graph embedding models, and each customer can be represented by a low-dimensional dense vector (i.e., embedding) for further segmentation analysis in Stage 3.

In Stage 3, the obtained embeddings of customers are clustered by the K-means algorithm into different groups, i.e., customer segments. Then, the clustering results are compared between different constructed networks and embeddings to select the best customer segments. Finally, customer needs for different customer segments are analyzed and discussed using product attribute analysis techniques such as IPA. Manufacturers can design new products or improve existing ones according to the customer needs of targeted customer groups. The proposed framework enables efficient and systematic data processing and analysis for customer segmentation and needs analysis from online reviews. More details and key techniques involved in each stage are provided in the following subsections.

3.2 Data Processing and Information Extraction

3.2.1 Data Preprocessing. Customer online reviews on targeted products can be gathered from e-shopping websites and online forums by web crawling techniques. These reviews usually include customers' opinions and use experience on products, overall product ratings, and order information such as purchase location and purpose. Since the collected raw data is often unstructured, data cleaning and preprocessing are needed. Commonly used procedures include data cleaning (e.g., removing redundant, irrelevant, erroneous, duplicate data and unnecessary symbols, populating or eliminating missing values), sentence segmentation, and word tokenization. The operations mentioned here are a non-exhaustive list, and readers should choose appropriate ones according to the actual situation. For example, the raw text of a typical customer review on automobiles crawled from the Internet looks like "*Word of mouth; &nbsp; December 18, 2016, from Autohome Android version; [Fuel] The fuel consumption is quite high, but there is no major problem with the power and weight of the car …* ". After data cleaning, those redundant words (e.g., "*Word of mouth*"), irrelevant words (e.g., "*&nbsp*", "*from Autohome Android version*") and meaningless symbols (e.g., *[]*) will be removed. Then, the clean data will be the input to the next steps to support the generation of product attribute vectors and customer vectors, whose processes are detailed in Secs. 3.2.2 and 3.2.3, respectively.

3.2.2 Identification of Product Attributes. After getting clean online reviews, the major product attributes mentioned by customers should be extracted. The common TF-IDF (Term Frequency–Inverse Document Frequency) algorithm [35] is used to screen out the high-frequency keywords from online reviews. Since TF-IDF assumes that the importance of a word is proportional to the number of times it appears in a document of interest (e.g., a piece of online review), but inversely proportional to the frequency of its occurrence in the whole corpus, it prevents most high-frequency words irrelevant to our research domain (e.g., user experience and opinions on motor vehicle) from being selected. Then, these keywords are transformed into word vectors with the Skip-Gram model [36], which not only can transform text into easily processed numerical values but is also good at digging potential semantic relationships of the words in the text. Finally, these word vectors are grouped into clusters using the X-means method [37]. Other popular method such as the LDA model is also tested but its performance is poor since one keyword can be often assigned to multiple topics/attributes, which fails to satisfy the requirement of accurate mapping between keywords and attributes.

We manually name each cluster according to the functional structure of the product (e.g., a motor vehicle usually consists of systems in power, chassis, body, and electronics), and the name of a cluster

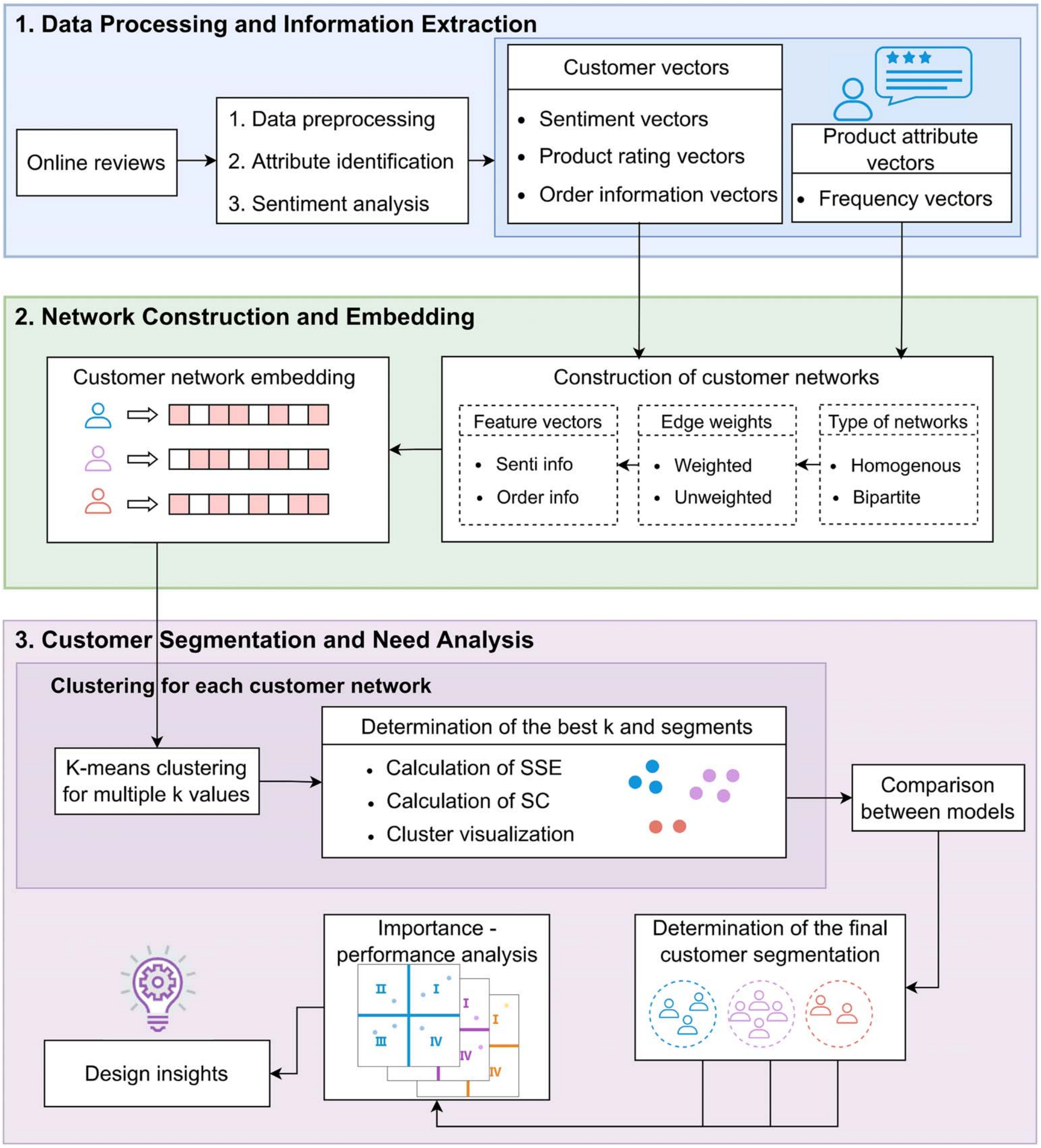

1. Data Processing and Information Extraction
Online reviews
1. Data preprocessing
2. Attribute identification
3. Sentiment analysis
Customer vectors
Sentiment vectors
Product rating vectors
Order information vectors
Product attribute vectors
Frequency vectors
2. Network Construction and Embedding
Customer network embedding
Construction of customer networks
Feature vectors
Senti info
Order info
Edge weights
Weighted
Unweighted
Type of networks
Homogenous
Bipartite
3. Customer Segmentation and Need Analysis
Clustering for each customer network
K-means clustering for multiple k values
Determination of the best k and segments
Calculation of SSE
Calculation of SC
Cluster visualization
Comparison between models
Importance - performance analysis
II
I
III
IV
Determination of the final customer segmentation
Design insights

Fig. 1 The overall framework for customer segmentation and need analysis

is an identified product attribute. Then, we further refine the obtained product attributes, including merging those attributes sharing the same product function (e.g., horsepower and acceleration both belong to the Power attribute). We also manually filtered out those keywords irrelevant to any product attributes according to the functional structure or evaluation of the product to ensure the correctness of extracted keywords in each cluster.

The name of each cluster (i.e., identified product attribute) and the significant words (i.e., keywords obtained from screening with TF-IDF and manual filtering) in each cluster can then form an attribute-keyword dictionary, and its detailed establishment process can be found in our previous work [38]. This dictionary can be used to quickly identify those frequently mentioned product attributes from new online reviews. Note that the one keyword can only be found in one cluster (i.e., one product attribute). Thus, once a keyword from the attribute-keyword dictionary is detected from the text of an online review, the mentioned product attribute can be identified.

Specifically, these new reviews are first segmented into words and phrases based on commas, semicolons, and other punctuation marks, and part-of-speech tagging is applied to get the corresponding lexicality tags (e.g., nouns, verbs, etc.) of these words and phrases. The nouns or noun phrases are then searched in the attribute-keyword dictionary to identify what product attributes are mentioned. The occurrence frequencies of the keywords in online reviews are calculated as the product attribute vectors as shown in the following equation:

$$P_j = [p_{j1}, p_{j2}, \ldots, p_{jk}] \tag{1}$$

Here, P_j is the product attribute vector for product attribute j, and k is the number of keywords for the product attribute, and p_{jk} is the occurrence frequency of the keyword k for attribute j.

3.2.3 Sentiment Analysis. After obtaining the identified product attributes, we also need to get customers' sentiments toward these attributes. A tricky problem here is that customers sometimes express their opinions on one product attribute in multiple sentences (i.e., a customer may mention one product attribute multiple times), and sometimes, they mention multiple product attributes in one sentence. If we perform sentiment analysis simply by the sentences separated by punctuations, the bias is inevitable since the sentiment analysis model may not be able to pair customers' opinions with corresponding product attributes accurately. To overcome this issue, we develop a semantic group-based method. A semantic group is a collection of descriptions of a single product attribute from one customer's product review, and one customer's review can include one or more semantic groups. In other words, a semantic group is an attribute–description pair. For example, the sentence "*This phone's battery is very durable*" can converted into one semantic group ("battery", "very durable"), while the sentence "*The fuel consumption and power are surprisingly good for this vehicle*" can be converted into two semantic groups, ("fuel," "surprisingly good") and ("power," "surprisingly good"). If a customer mentions one product attribute multiple times, those generated semantic groups sharing the same attribute will be merged. When counting the identified product attributes and calculating sentiment scores, semantic groups can be used as the fundamental textual units, and customers' attitudes toward products can be more accurately captured. The generation of semantic groups from customer reviews includes sentence segmentation, keyword matching, and dependency parsing analysis. For more details on this method, please refer to our previous work [39].

The obtained semantic groups are then put into sentiment analysis models to get their sentiment scores. We found that the sentiment classification probabilities generated by different methods are often inconsistent (sometimes they even contradict each other). It may not be appropriate to directly use the classification probabilities from either of these sentiment analysis models or their averages as the final sentiment scores. Therefore, to reduce potential biases and improve the model performance, we adopt a voting mechanism for sentiment analysis using two pre-trained models: the ERNIE (Enhanced representation through knowledge integration) model [40] and the BiLSTM (Bidirectional Long Short-Term Memory) model [41]. ERNIE can accurately classify sentiments by learning the language representation and semantic knowledge at the entity and phrase levels in sentences. BiLSTM realizes sentiment analysis by capturing bidirectional semantic dependencies and contextual information. For each semantic group, both models will generate two probabilities within the range of [0, 1]. One probability is about the positive sentiment, and the other is about the negative sentiment. These probabilities are also called sentiment intensity values, and a larger value represents stronger sentiment intensity. The final sentiment polarity (i.e., positive, negative, or neutral) of a semantic group is determined by the voting mechanism based on the sentiment intensity values obtained from two models using the following rules:

$$\theta = \begin{cases} 1, & \text{if } Pos_1 \geq c_0 \text{ and } Pos_2 \geq c_0 \\ -1, & \text{if } Neg_1 \geq c_0 \text{ and } Neg_2 \geq c_0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Here, θ is the final sentiment polarity, which is converted to numerical values (i.e., "positive" to 1, "neutral" to 0, and "negative" to −1). c_0 is the cutoff point. Pos_1 and Neg_1 are the positive and negative sentiment intensity values generated from the BiLSTM, Pos_2 and Neg_2 are generated from the ERNIE model.

We use two models to eliminate possible bias in sentiment analysis since different corpus have been used in their respective training processes (i.e., these models have been trained on other datasets). We tested different cutoff points c for the above classification and found that the accuracy of sentiment analysis is the largest (about 93%) when the cutoff point is set to 0.8. Here the accuracy is defined as the percentage of the number of correct sentiment predictions over the total number of predictions. We manually labeled the sentiment polarities for each attribute from 5870 online reviews as the ground truth. Other alternative method such as the supervised BERT for sentence pair classification (BERT-SPC) model is also tested with a classification accuracy of 71.78%. Without training with textual materials such as reviews from Chinese automobile forums, the performance of supervised learning methods like BERT-SPC is limited in our research context.

After obtaining customers' sentiments, customer sentiment vectors can be generated using his/her sentiment polarities on each product attribute. Then, customer sentiment vectors, product rating vectors, and order information vectors can be concatenated into customer vectors to represent customer characteristics as shown in Eq. (3):

$$S_i = \text{CONCAT}([\theta_{i1}, \theta_{i2}, \ldots \theta_{ia}], [r_{i1}, r_{i2}, \ldots, r_{ib}], [o_{i1}, o_{i2}, \ldots o_{ic}]) \tag{3}$$

Here, S_i is the customer vector for customer i. θ, r, and o are the sentiment polarity, product rating and value of order information, respectively. a, b, and c are the number of product attributes, dimension of product rating, and length of the order information vectors, respectively.

3.3 Network Construction and Embedding.

After obtaining the customer and product attribute vectors from online review data, a social network of customers can be established. Since the network embedding can reduce the dimension of the high-dimensional customer vectors and capture the complex interrelationships between customers, network embeddings will be generated to better represent customers. The following two subsections provide network construction and embedding processes for homogenous networks and bipartite networks in detail.

3.3.1 Homogeneous Network Construction and Embedding.

A homogeneous customer social network, $G(U, E)$ is constructed to capture the complex and latent interrelationships and nuances among customers. Here, U is a set of nodes representing customers, and E denotes a set of links representing the connections among customers. If the cosine similarity (a classical metric to measure the similarity of vectors [42], see Eq. (4)) of the customer vectors of node i and node j exceeds a threshold of α, a link $E_{i,j}$ between these two nodes is formed (see Eq. (5))

$$\text{Sim}(i, j) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}| \times |\vec{S_j}|} \tag{4}$$

$$E_{i,j} = \begin{cases} 1, & \text{Sim}(i, j) \geq \alpha \\ 0, & \text{else} \end{cases} \tag{5}$$

Here, S_i and S_j are the customer vectors corresponding to customer i and j. The threshold of similarity α is usually selected as 0.5 according to previous studies [43]. In the constructed social network of customers, two customers are more likely to be linked if they make comments on similar product attributes with comparable sentiment polarities, which implies they may have similar customer needs. The customer vectors can include partial or all available customer characteristics. Therefore, different customer networks can be constructed and compared to obtain better clustering results.

In our study, GAE is leveraged to learn the embeddings of the nodes in the constructed social network of customers due to its strong performance in capturing hidden features in complex network structures [44]. Figure 2 shows the overall structure of GAE. The input of GAE includes the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ of the constructed social network and the feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times F}$ of the nodes, in which N is the number of nodes and F is the dimension of customer vector. The encoder of the GAE is a multilayer graph convolutional network given by Eq. (6) [45]:

$$\boldsymbol{Z} = \text{GCN}(\boldsymbol{X}, \boldsymbol{A}) \tag{6}$$

$\boldsymbol{Z} \in \mathbb{R}^{N \times S}$ is the output of the encoder, a matrix formed by embedded vectors of nodes, and S is the dimension of the embedded vector. GCN(:) represents a graph convolutional neural network,

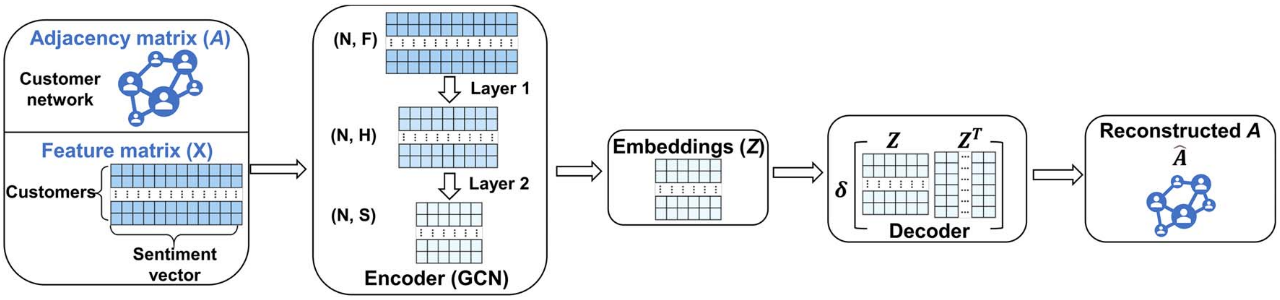


Fig. 2 Overall structure of the Graph Autoencoder (GAE) for homogenous network embedding

which consists of two graph convolutional layers as shown in the encoder part of Fig. 2. No activation function is added to the second graph convolutional layer so that the embedded information can be fully expressed. The form of graph convolution is shown in Eq. (7):

$$\mathrm{GCN}(\boldsymbol{X}, \boldsymbol{A}) = \tilde{\boldsymbol{A}}\mathrm{ReLU}(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_0)\boldsymbol{W}_1 \tag{7}$$

$\boldsymbol{W}_0 \in \mathbb{R}^{F\times H}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{H\times S}$ are the parameter matrices to be learned in the first and second graph convolutional layers (i.e., layer 1 and layer 2), respectively. H and S represent the dimensions of the first and second layers. Normalized Laplacian matrix is calculated as Eq. (8):

$$\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{1/2}(\boldsymbol{A} + \boldsymbol{I}_N)\tilde{\boldsymbol{D}}^{1/2} \tag{8}$$

where $\boldsymbol{I}_N$ is the identity matrix. The matrix $\tilde{\boldsymbol{D}}$ is a diagonal matrix of nodal degrees. The diagonal element of this matrix is shown in Eq. (9):

$$\tilde{\boldsymbol{D}}_{ii} = \sum_{j=1}^{N}(\boldsymbol{A} + \boldsymbol{I}_N)_{ij} \tag{9}$$

where i, j are the indices of a matrix element.

The decoder of GAE reconstructs the network so that the implicit features of the data can be learned, and it consists of a matrix multiplication between latent variables with an activation function. The output of GAE is a reduced graph adjacency matrix $\hat{\boldsymbol{A}}$ as shown in Eq. (10):

$$\hat{\boldsymbol{A}} = \delta(\boldsymbol{Z}\boldsymbol{Z}^{\boldsymbol{T}}) \tag{10}$$

where $\boldsymbol{Z}$ is the low-dimensional vector representation of the nodes obtained by the encoder, and δ is a logistic sigmoid function. $\boldsymbol{Z}$ should make the reconstructed adjacency matrix $\hat{\boldsymbol{A}}$ as similar to the original adjacency matrix $\boldsymbol{A}$ as possible since the adjacency matrix determines the structure of the network. As a result, cross-entropy is adopted as the loss function in the model training as seen in Eq. (11).

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y\log\hat{y} + (1-y)\log(1-\hat{y}) \tag{11}$$

where y is the value of the element (i.e., 0 or 1) in the original adjacency matrix. $\hat{y}$ is the value of the corresponding element in the reconstructed adjacency matrix. In a word, GAE embeds the customer vectors into low-dimensional dense vectors by minimizing the reconstruction error and keeping as much structural information of customers' social network as possible. These embeddings are then used in the further clustering of customers.

3.3.2 Bipartite Network Construction and Embedding. When considering the interactions between customers and product attributes, a bipartite network can be constructed. In a bipartite network (U, V, E), U represent customer nodes, V represent product attribute nodes, and E represent the connection between customers and product attributes (e.g., customer sentiments to product attributes). Customer nodes and product attribute nodes are linked when the customer mentions the attribute in online reviews, and the edge weight (represented by S) is determined by the sentiment polarities.

In this study, bipartite GraphSAGE [28,46] is utilized for the bipartite network embedding considering its efficiency for processing large-scale graphs. It can generate low-dimensional embeddings by sampling and aggregating information from local neighbor nodes. Figure 3 shows the structure of bipartite GraphSAGE. The input contains the features of customers $X_u = \{x_u, \forall u \in U\}$ and features for product attributes $X_v = \{x_v, \forall v \in V\}$. Here, $X_u \in \mathbb{R}^{N\times d_u}$ and $X_v \in \mathbb{R}^{M\times d_v}$. The embeddings of the nodes are initially set by the node features. Then, the model iteratively updates the embeddings of customer nodes by aggregating the local neighbor information for product attribute nodes as shown in Eq. (12):

$$h^p_{N(u)} \leftarrow M^u_v \cdot \mathrm{AGGREGATE}^p_u(\{h^{p-1}_v, \forall v \in N(u)\}) \tag{12}$$

where $N(u)$ is the immediate neighborhood of the customer node.

Similarly, the embeddings of the product attribute nodes are the aggregated neighbor customer embeddings:

$$h^p_{N(v)} \leftarrow M^v_u \cdot \mathrm{AGGREGATE}^p_v(\{h^{p-1}_u, \forall u \in N(v)\}) \tag{13}$$

Here $h^p_{N(u)} \in \mathbb{R}^{d^{p-1}_u}$ and $h^p_{N(v)} \in \mathbb{R}^{d^{p-1}_v}$ are the aggregations of the neighbor nodes in step p. h^{p-1}_u and h^{p-1}_v are the embeddings of customers and product attributes in step p-1. $M^u_v \in \mathbb{R}^{d^{p-1}_u\times d^{p-1}_v}$ and $M^v_u \in \mathbb{R}^{d^{p-1}_v\times d^{p-1}_u}$ are the transformation matrices from product attribute to customer and from customer to product attribute, respectively. The mean aggregator is selected to aggregate the information of neighbor nodes by taking the average.

After the aggregation, the embeddings can be updated by the following equations:

$$h^p_u \leftarrow \sigma(W^p_u \cdot \mathrm{CONCAT}(h^{p-1}_u, h^p_{N(u)})) \tag{14}$$

$$h^p_v \leftarrow \sigma(W^p_v \cdot \mathrm{CONCAT}(h^{p-1}_v, h^p_{N(v)})) \tag{15}$$

Here $h^p_u \in \mathbb{R}^{d^p_u}$ and $h^p_v \in \mathbb{R}^{d^p_v}$ are the updated embeddings in step p. $W^p_u \in \mathbb{R}^{d^p_u\times 2d^{p-1}_u}$ and $W^p_v \in \mathbb{R}^{d^p_v\times 2d^{p-1}_v}$ are the weight matrices for customers and product attributes. At final step P, we can obtain the final embedding output as $z_u \equiv h^P_u$ and $z_v \equiv h^P_v$ for customers and product attributes, respectively ($z_u \in \mathbb{R}^{d^P_u}$,$z_v \in \mathbb{R}^{d^P_v}$). In our study, both customer vectors and product attribute vectors are row vectors. Thus, the concatenation is a horizontal operation. As shown in Eq. (16), the bipartite graph-based loss function is utilized to train the model and learn the parameters to ensure the embeddings of the nodes are similar to their neighbors.

$$\begin{aligned} J_{\mathrm{BG}} = &-\log\,[\sigma(f[\mathrm{CONCAT}(z_u, z_v), S((u, v))])] \\ &-Q_u \cdot E_{u_n\sim P_n(u)}\log[\sigma(f[\mathrm{CONCAT}(z_{u_n}, z_v), \gamma])] \\ &-Q_v \cdot E_{v_n\sim P_n(v)}\log[\sigma(f[\mathrm{CONCAT}(z_u, z_{v_n}), \gamma])] \end{aligned} \tag{16}$$

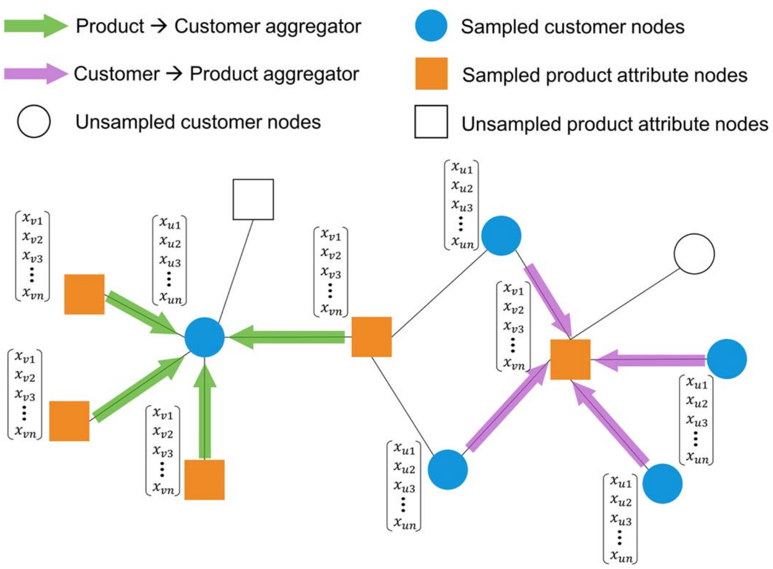

Product → Customer aggregator
Customer → Product aggregator
Unsampled customer nodes
Sampled customer nodes
Sampled product attribute nodes
Unsampled product attribute nodes
x_{v1} x_{v2} x_{v3} ⋮ x_{vn}
x_{u1} x_{u2} x_{u3} ⋮ x_{un}
x_{v1} x_{v2} x_{v3} ⋮ x_{vn}
x_{u1} x_{u2} x_{u3} ⋮ x_{un}
x_{v1} x_{v2} x_{v3} ⋮ x_{vn}
x_{v1} x_{v2} x_{v3} ⋮ x_{vn}
x_{v1} x_{v2} x_{v3} ⋮ x_{vn}
x_{u1} x_{u2} x_{u3} ⋮ x_{un}
x_{u1} x_{u2} x_{u3} ⋮ x_{un}
x_{u1} x_{u2} x_{u3} ⋮ x_{un}

Fig. 3 Overall structure of bipartite GraphSAGE

Here, (u, v) is the edge between the customer and product attribute node, and $S((u, v))$ is the weight of the edge. f is a fully connected neural network to generate the similarity between the concatenation of customer node embedding and product attribute node embedding (CONCAT(z_u, z_v)), and the corresponding edge weight ($S((u, v))$). Q_u and Q_v are the number of negative samples for customers and product attributes. Here, negative samples represent the edges that do not really exist in the network (e.g., in a 3-node network, only edges between nodes 1 and 2, 2 and 3 exist. Then, the non-existing edge between nodes 1 and 3 can be considered as a negative sample). Including negative samples in the loss function can balance the training data and make the model more robust. σ is the sigmoid function. P_n is the negative sampling distribution, and γ is the weight of negative samples.

3.4 Customer Segmentation and Need Analysis

3.4.1 Customer Segmentation Based on Clustering of Customer Node Embeddings. After the embeddings of customer vectors are obtained, clustering of these embeddings can be performed. In this study, k-means [47] is selected as the clustering method, which can divide customers into k groups. The main reason for choosing k-means is that the number of clusters is controllable. Designers and manufacturers can select the number of clusters actively and compare the performance according to their preferences, expectations, and even some commercial considerations. Although alternative methods such as X-means can automatically determine the number of clusters, if the number of customer segments generated from X-means is too small or too large, designers will struggle to interpret the practical meanings of these segments and develop associated design strategies. To choose the appropriate k value, the common elbow method is utilized for its simplicity and effectiveness in identifying the optional number of clusters. For each number of clustered groups k (2,3,4,5…10), the sum of squared distance (SSD) from each point to the cluster center is calculated as shown in Eq. (17):

$$\mathrm{SDD} = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{17}$$

Here, p represents the center of cluster C_i, m_i is a point in the cluster C_i, and k is the number of groups to be clustered. A smaller SSD means the data points in each cluster are closer to the cluster center. Then, these SSD values are plotted against the number of groups k. In the generated curve, the point with the most significant curvature change is selected as the elbow point, indicating the increase of number for clusters cannot significantly improve the clustering results. The appropriate range of k is selected around the corresponding number of clusters of the elbow point. The clustering results of each appropriate k in the range are visualized with t-Distributed Stochastic Neighbor Embedding (t-SNE) [48]. t-SNE is commonly used to visualize high-dimensional data in low-dimensional space. It realizes dimensional reduction by calculating the joint probabilities between data points, and minimize the Kullback-–Leibler (KL) divergence between the probabilities of the high-dimensional and low-dimensional data. Usually, the SSD value decreases with the increase of the cluster number. The final k value is determined based on both SSD and the visualization results.

According to our previous tests, we recommend the following rules for evaluating the visualization results. First, customers should be closely aggregated within a cluster. Second, the clusters need to have clear boundaries, and the customers of different groups should not be mixed together. Finally, the sizes of the different generated clusters should be comparable, which prevents a single customer segment from dominating the overall clustering results.

To examine the clustering effect, we leverage a clustering evaluation index, silhouette coefficient (SC) [49], which combines the evaluation of the cohesion effect and separation effect of clustering. Compared to SSD that is used to evaluate clustering performance within the same method, SC is more suitable for comparing the clustering results between different methods. The value of SC is in the range of [− 1, 1]. The closer its value to 1, the better the clustering performance. The calculation of the silhouette coefficient is shown in Eqs. (18) and (19):

$$si = \frac{d_b(i) - d_a(i)}{\max\,(d_a(i), d_b(i))} \tag{18}$$

$$\mathrm{SC} = \frac{\sum_{i=1}^{N} s(i)}{N} \tag{19}$$

where $s(i)$ is the silhouette coefficient of a clustered point i, $d_a(i)$ is the average distance between point i and other points in the same cluster, and $d_b(i)$ is the average distance between point i and other points in different clusters. The overall silhouette coefficient SC of clustering is the average of the silhouette coefficients of all N clustered points. By comparing the visualizations and the SC values, the final segmentation can be determined.

3.4.2 Importance and Performance Analysis. After obtaining the customer segments, product attribute analysis techniques such as IPA can be applied to each customer segment. The performance is calculated using the average sentiment scores for the attribute. We use the extreme gradient boosting (XGBoost) model to estimate the influence of customer sentiments on the overall customer ratings of the product, and the importance is evaluated with the gain-based feature importance for the model. Finally, the IPA results can be plotted for the need analysis of each customer segment. The detailed procedures and calculations can be found in our previous study [39].

4 A Case Study of Passenger Vehicle

To demonstrate the proposed framework, we conduct a case study by utilizing online reviews of a passenger vehicle in China's automotive market. The proposed framework is implemented in PYTHON on a computer with Intel i7-10700F CPU and 16 GB RAM. The network construction process takes about 30 min and the graph embedding costs around 3 min for homogenous networks, and 5 min and 4 min for bipartite networks. The time of network construction is mainly influenced by the determination mechanism of network edges. For a homogenous network, the similarity of each pair of customers needs to be calculated. For bipartite networks, we only need to determine if a customer mentions one product attribute. Thus, the network construction for bipartite networks can be much faster than for homogenous ones. In the following subsections, the dataset used, the results of product attributes identification and sentiment analysis, customer network construction and segmentation, and need analysis of customer segments are presented and discussed.

4.1 Description of the Dataset. We collected 2986 valid reviews on Buick Envision (a typical midsize sport utility vehicle (SUV) with high-sale volume in China) posted between November 2016 and November 2021 from a popular Chinese auto forum [50] using web crawling techniques. Customers' reviews include their user experience on the vehicle and ratings on eight dimensions of the vehicle (space, power, fuel consumption, comfort, appearance, interior design, value for money, and operability). Also, the reviews contain order information such as the price, location, and purpose of purchase. One typical review includes around 800–1000 words. The collected reviews are first cleaned and preprocessed as explained in Sec. 3.2.1. All aforementioned preprocessing operations such as data cleaning, sentence segmentation, and word tokenization have been actually employed in the case study, and all personal information is removed. The sentences and words are segmented with the "Jieba" and "LTP" libraries in PYTHON [51].

This dataset provides real buyers' experience and opinions on a popular passenger vehicle from the largest automotive market in the world across 5 years. Researchers in Engineering Design can utilize the rich information of this dataset in multiple areas, such as product positioning, customer segmentation, feature enhancement, and design trend analysis. Our dataset can facilitate novel research topics such as comparing the customers' thoughts before and after COVID-19, between conventional vehicles and electric vehicles, between China and other developing or developed countries, for investigating the evolution of product design strategies in time, technology, and space domains. Researchers can also employ this dataset to develop techniques for recognizing synthetic online reviews and get safer design insights by excluding the increasing impact of Artificial Intelligence Generated Content (AIGC) in recent years. The dataset has been uploaded to GitHub[3] for open access, and a note documentation is also provided to guide users in accessing and utilizing the data effectively. The original language of this dataset is Chinese, and we also provide an English version (translated) for more convenient use by the public.

4.2 Results of Product Attributes Identification and Sentiment Analysis. The keywords are extracted from the preprocessed data and the attribute-keywords dictionary is established as described in Sec. 3.2.2. Based on the functional components of the vehicle system, 28 product attributes are finally identified as shown in Table 1. According to the system structure of the vehicle and customer experience, these attributes are further classified into five major categories: Engine (A_1), Chassis (A_2), Electrical and Control (A_3), Body (A_4), and Maintenance and Comfort (A_5). The 28 identified product attributes are associated with 472 keywords, and the average number of keywords under each attribute is 16.86.

Then, the semantic groups are generated using the identified product attributes, and the sentiment score of each semantic group can be calculated using the sentiment analysis models (ERNIE and BiLSTM). According to the classification rules introduced in Sec. 3.1.3, each semantic group is assigned with corresponding sentiment polarity label (positive, neutral, and negative). Then, the sentiment vectors of customers can be formed. Table 2 provides samples of the sentiment analysis results. For example, review 1 shows positive opinions on both vehicle engine systems and controllability, while review 2986 expresses neutral and negative opinions on these two attributes. The dimension of a customer sentiment vector is 84, since there are three sentiment polarity labels for each of the 28 identified product attributes from A_{11} (engine system) to A_{54} (controllability).

4.3 Results of Network Construction, Embedding, and Segmentation. As explained in Sec. 3.3, customer networks can be constructed with different network architectures such as different types of networks and different customer feature vectors. In this study, the order information vectors contain the informative customer characteristics such as rating, purchase location, and purchase purpose. Other information such as purchase price and fuel consumption is not considered since only one type of product is considered. After the construction, network embedding models are applied to obtain low-dimensional customer vectors. Finally, the vectors are clustered with the k-means clustering method. The clustering results are compared based on the silhouette coefficient and the visualizations. We set the clustering results without network construction and embedding as the alternative methods. The traditional community detection method is also applied for the method comparison. Table 3 shows the nine models under our framework tested in this study, containing both the homogeneous networks and bipartite networks. In addition, we also expect to compare with two alternative methods from existing literature [8,9]. The results for the construction, segmentation, and comparison are discussed in detail as follows.

4.3.1 Results of Homogeneous Networks Construction, Embedding, and Segmentation. Since customer characteristics are mainly reflected in customer sentiment vectors, for homogeneous networks, a sentiment network of customers is first constructed. If the cosine similarity of two customer sentiment vectors exceeds a threshold, these two customers can form a link. In our study, the threshold value is set to 0.5, which is a commonly used value in social network research [52]. In addition, this threshold value can make the degree of customer nodes larger to retain more structural information about the network. Figure 4 presents part of the constructed social network of customers. Here the red and green nodes represent the customers whose reviews focus more on Body (A_4) or Maintenance and Comfort (A_5) of the vehicle, respectively. We can see that customer nodes with the same color are more likely to be linked.

The constructed network is embedded using the GAE model (i.e., Model 1), and the dimension of generated dense vectors has been reduced from 84 to 10. There is no straightforward guideline for

[3] https://github.com/wobuhuiya/Online-Review-Analysis/tree/main

Table 1 Five major categories and 28 product attributes identified from online reviews

Major category	Product attributes
Engine ($\boldsymbol{A}_1$)	Engine system (A_{11}), Valve system (A_{12}), Engine room (A_{13}), Power (A_{14})
Chassis ($\boldsymbol{A}_2$)	Fuel system (A_{21}), Fuel consumption (A_{22}), Transmission (A_{23}), Braking (A_{24}), Steering (A_{25}), Driving (A_{26}), Suspension (A_{27})
Electrical and Control ($\boldsymbol{A}_3$)	Circuit device (A_{31}), Anti-theft system (A_{32}), Electronic control (A_{33}), Meter (A_{34}), Lamps (A_{35}), Air conditioner (A_{36}), Switch (A_{37}), Multimedia (A_{38})
Body ($\boldsymbol{A}_4$)	Body accessories (A_{41}), Interior accessories (A_{42}), Body space (A_{43}), Car door (A_{44}), Body design (A_{45})
Maintenance and Comfort ($\boldsymbol{A}_5$)	Maintenance (A_{51}), Interior (A_{52}), Comfort (A_{53}), Controllability (A_{54})

Table 2 Sample results of the sentiment analysis on customers' reviews

	A_{11j} (Engine system)			…	A_{54j} (Controllability)		
Review ID (j)	A_{11j}^{pos}	A_{11j}^{neu}	A_{11j}^{neg}	…	A_{54j}^{pos}	A_{54j}^{neu}	A_{54j}^{neg}
1	1	0	0	…	1	0	0
2	0	0	0		0	0	0
…	…	…	..	…	…	…	…
2986	0	1	0	…	0	0	1

selecting the embedding dimension S. Usually, S is determined by try and error by examining the clustering performance (e.g., SC as mentioned in Sec. 3.4) for each S chosen. These embeddings are then clustered into several groups (i.e., customer segments) using the k-means method.

Figure 5 shows the plot for the change of SSD and SC with the number of clusters k. From the plot of SSD, when $k = 4$, there is a noticeable change in the rate of decrease, forming an "elbow." The number of segments is selected around this elbow point at $k = 4$. Then, the plot of SC shows that the value of SC at $k = 4$ and $k = 5$ is similar, while the value of SSD is smaller for $k = 5$. Therefore, we determine the final k between these two points by the visualization of the clusters.

Figure 6 demonstrates the visualizations generated using the t-SNE model. In this figure, each dot represents a customer, and the color indicates the corresponding segment (i.e., cluster) of this customer. We can see that the clustered customers generated from the proposed approach show clear boundaries between different groups. For $k = 5$, the results for the first four clusters are similar to those for $k = 4$. However, the number of customers in the fifth segment (i.e., Segment 4) is too small compared with other segments. Then, it may lead to excessive and unnecessary costs to analyze the customer needs in this segment and customize the products. Thus, for the sentiment network, we generate four customer segments with the value of SC equals to 0.53.

To obtain good performance for the clustering, we test different network architectures. Besides customer sentiments, other order information from online reviews can also be included in customer vectors to represent customer characteristics. The available information includes the car model, purchase price, time, purpose, location, fuel consumption, and ratings. Since the collected online reviews are for the same vehicle model, we only consider the purchase purpose, location, and ratings which reflect distinguishable customer characteristics. The data are preprocessed with data standardization and vectorization. Then, the generated vectors are concatenated with customer sentiment vectors as the customer feature vectors. The process of network construction and segmentation is similar to that in Sec. 4.2.1. The clustering results from this model (i.e., Model 2) are plotted in Fig. 7(*a*).

In Fig. 7(*a*), the boundaries between segments are still clear, but the data in the clusters are not as dense as the clusters generated with the customer sentiment vectors. Also, the calculated silhouette coefficient is smaller. Although more customer information is considered in the network construction, the performance of clustering is worse. The reason may be that the purchase purpose and location have minimal impact on customers' evaluation of the products, and customer ratings can be inferred from customer sentiments. The addition of this information makes the distinction between customers blurred. Therefore, the performance of this network architecture is poor in our case study, but it may work for the data that contains rich and clear customer personalities.

In order to examine the performance of using customer sentiments and network embedding, we compare the clustering results for another two alternative methods. In the first alternative method, instead of using GAE for graph embedding, we directly use k-means to cluster the customer sentiment vectors with a dimension of 84 (i.e., Model 3). In the second alternative method, the customer vectors only describe whether a product attribute is mentioned or not based on the generated semantic groups, and no sentiment information is included (i.e., Model 4). Thus, the dimension of these vectors is 28, and these vectors are defined as the customer mention vectors. Then, the customer vectors are embedded using the customer network and clustered into customer segments. The purpose of comparing with these alternative methods is to illustrate the necessity of incorporating sentimental information into clustering and the effectiveness of graph

Table 3 Nine different models for the constructed networks and the segmentation methods

Model ID	Network type	Sentiment info	Order info	With edge weight	Embedding method	Clustering method
1	Homogeneous	Yes	No	No	GAE	k-Means
2	Homogeneous	Yes	Yes	No	GAE	k-Means
3	—	Yes	No	—	—	k-Means
4	Homogeneous	No	No	No	GAE	k-Means
5	Homogeneous	Yes	No	No	—	Community detection
6	Bipartite	Yes	Yes	Yes	GraphSAGE	k-Means
7	Bipartite	No	Yes	Yes	GraphSAGE	k-Means
8	Bipartite	Yes	Yes	No	GraphSAGE	k-Means
9	Bipartite	No	Yes	No	GraphSAGE	k-Means
Alternative method 1 [8]	—	Yes	No	—	—	X-Means
Alternative method 2 [9]	Homogeneous	Yes	No	No	—	Modularity Clustering

Note: The sentiment and order information are selected to be included in the customer vectors or not for comparison. Two alternative methods from existing literature [8,9] are also listed.

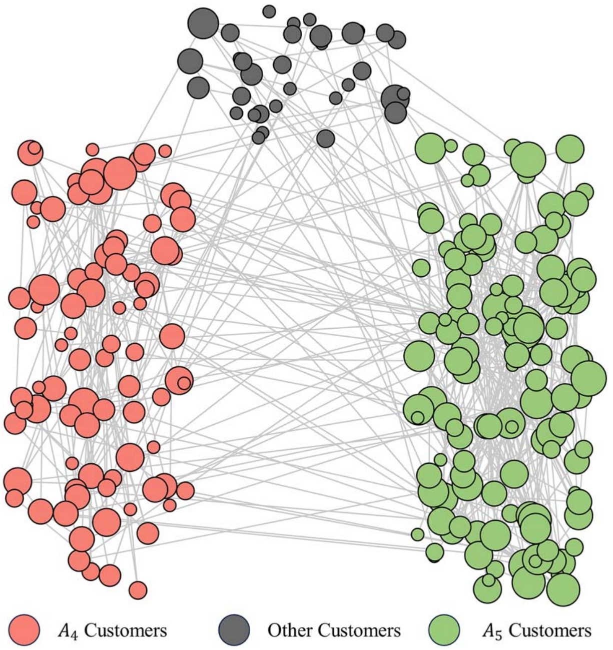


Fig. 4 Part of the constructed social network of customers. Red nodes represent customers who care most about body (A_4), while green nodes represent customers whose reviews focus on maintenance and comfort (A_5). This size of the node is proportional to its degree.

embedding for dimensionality reduction. The clustering results of Models 3 and 4 are shown in Figs. 7(*b*) and 7(*c*), respectively.

In Fig. 7(*b*), the boundaries between the clusters from the no-graph embedding methods are blurry. Also, the silhouette coefficient for the clustering results is small. Without graph embedding treatments, customer feature vectors are very sparse which makes the clustering more difficult. In Fig. 7(*c*) where the method uses no sentiment information, the segments are clearly clustered. However, compared with the results using the sentiment network and GAE model (Model 1), the sizes of the clusters for these results are imbalanced. The percentage of customers for all four segments is calculated as 17.2%, 4.2%, 71.1%, and 7.5%. Although the silhouette coefficient and the visualization indicate the clustering results may be better than the results from other methods, the segments may have drawbacks in the real world for customer segmentation. Without customer sentiments, customers' characteristics are weakened leading most customers to fall into segment 2. The results may not reflect the heterogeneity of customers, and the need for customer segmentation may be questioned.

For the homogeneous networks, we compared our methods with the community detection model (i.e., Model 5). Community detection is a traditional method for segmenting networks. The customer sentiment network is applied in this model, and the segmentation results are shown in Fig. 7(*d*). The generated clusters are similar to those generated with no graph embedding, and the performance of this method is poor. The customers in different groups are mixed together, and the boundaries are blurry.

To further validate the effectiveness of our framework, we also tested the performance of two alternative customer segmentation methods from literature. One is from Suryadi and Kim's work [8], in which they extracted customer attributes from online reviews for laptop products. The resultant customer vector consists of product features with sentiment polarity (positive/negative) and then X-means clustering was conducted for customer segmentation. The other alternative method is from Park and Kim's work [9], where they extracted customer attributes from online review data and built a customer network based on these attributes and predefined networking rules. Then this network is partitioned by modularity clustering to realize customer segmentation. Both methods are representative in the field of customer segmentation based on mining online reviews. As shown in Figs. 7(*e*) and 7(*f*), the silhouette coefficients (SC) obtained from two alternative methods are 0.05 and 0.10, respectively, which are much worse than the performance of Model 1 under our framework (SC = 0.53). In addition, the generated clusters are mixed together and the boundaries are blurry in Fig. 7(*e*). In Fig. 7(*f*), the number of red dots (i.e., Group 0) is quite small compared to other clusters, which weakens the necessity of analyzing the customer needs in this segment. The comparison results indicate that both alternative methods perform worse than our framework. One possible explanation is that these methods did not consider the challenge of clustering high-dimensional data. In addition, the clustering techniques used in these alternative methods (i.e., *X*-means and Modularity Clustering) automatically calculate the best number of clusters; thus, they leave no more room for designers to control the final generated number of clusters. Designers will struggle to interpret the practical meanings of an uncontrolled number of segments and develop associated design strategies.

4.3.2 Results of Bipartite Networks Construction, Embedding, and Segmentation. Considering that bipartite networks have more complex network structures and may contain more information,

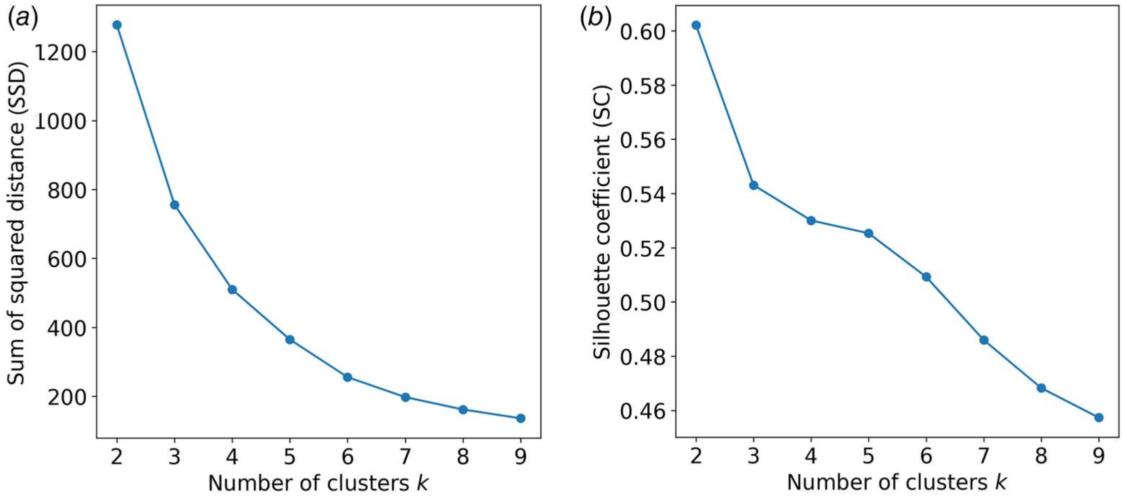


Fig. 5 The relationship between clustering performance and the number of clusters *k*: (*a*) the change of SSD and *k*, and (*b*) the change of SC and *k*. The smaller the SSD and larger the SC, the better.

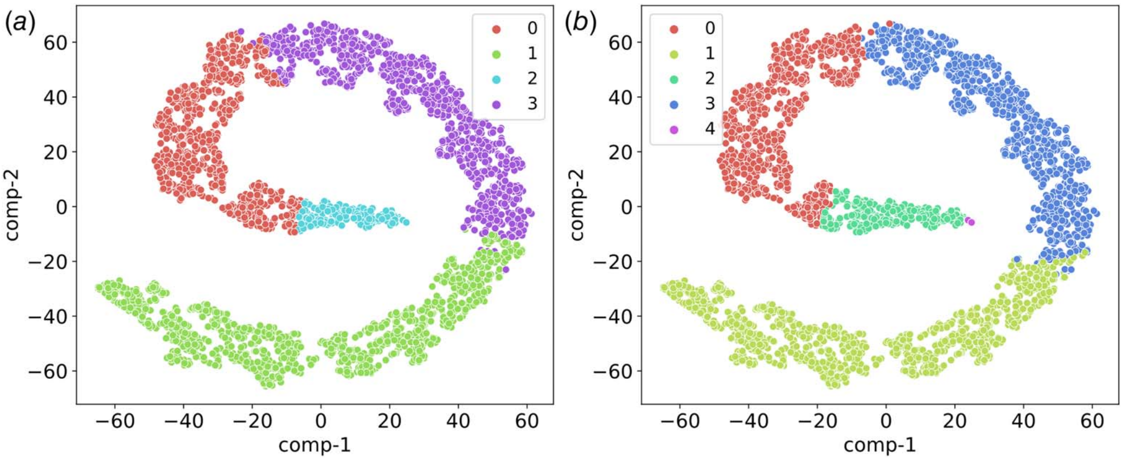


Fig. 6 Visualization of customer segments generated for customer sentiment networks (Model 1) with different number of clusters: (*a*) $k=4$, (*b*) $k=5$. Each color represents a segment.

we also construct these networks to search for better results for customer segmentation. Since the collected online reviews are for the same vehicle model, customers and product attributes are set as two types of nodes. The customer node and the product attribute node are linked if the customer mentioned the attribute in the review. The feature vectors of customer nodes are the concatenation of customer sentiment vectors and order information vectors. We also try to replace the customer sentiment vectors with the customer mention vectors and build different customer-attribute networks. For each attribute, the mentioned frequencies of the corresponding product keywords are calculated and combined as the feature vector. After the preprocessing, weighted networks and unweighted networks are constructed. For the weighted networks, customer sentiments are utilized as the weights. The edge weight is set as -1 for the negative polarity, 0 for the neutral polarity, and 1 for the positive polarity. Finally, four different customer-attribute bipartite networks (i.e., Models 6–9 in Table 3) are constructed with two choices for customer vectors and two choices for the edge weights. Figure 8 shows an example of part of the constructed bipartite network. Customers' characteristics can be found from their sentiments to their attributes. For example, customers 0–5 and customer 7 have average positive sentiments toward the product, while other customers have average neutral or even negative sentiments.

The constructed networks are then embedded into customer embeddings with the bipartite GraphSAGE model. The dimensions of generated dense vectors are also determined by trial and error. Similar to the process in Sec. 4.2.1, the embeddings are clustered by the *k*-means method and the segments are visualized using the t-SNE method. In this study, we compare the results from four different bipartite networks. The visualizations of the generated segments are shown in Fig. 9.

The first bipartite network in Fig. 9(*a*) is a weighted network with the customer sentiment information included in customer vectors (i.e., Model 6). It can be seen that the boundaries are clear but some customers for Segment 0 and Segment 3 are mixed. The corresponding silhouette coefficient is smaller than that in Sec. 4.2.1. Figure 9(*b*) is for the weighted network with the customer sentiment vectors replaced by the customer mention vectors (i.e., Model 7), and the performance of this model is slightly worse than the first bipartite network. The points in clusters are distributed looser.

The third network is an unweighted network, and the customer sentiment vectors are utilized as part of the feature vectors (i.e., Model 8). The clustering results are shown in Fig. 9(*c*). The results for this network model show the best SC value for bipartite networks. Compared with the first network, the information on customer sentiments is only considered once in the customer vectors but not in the edge weights. The avoiding duplication of sentiment information may be the reason for its better performance.

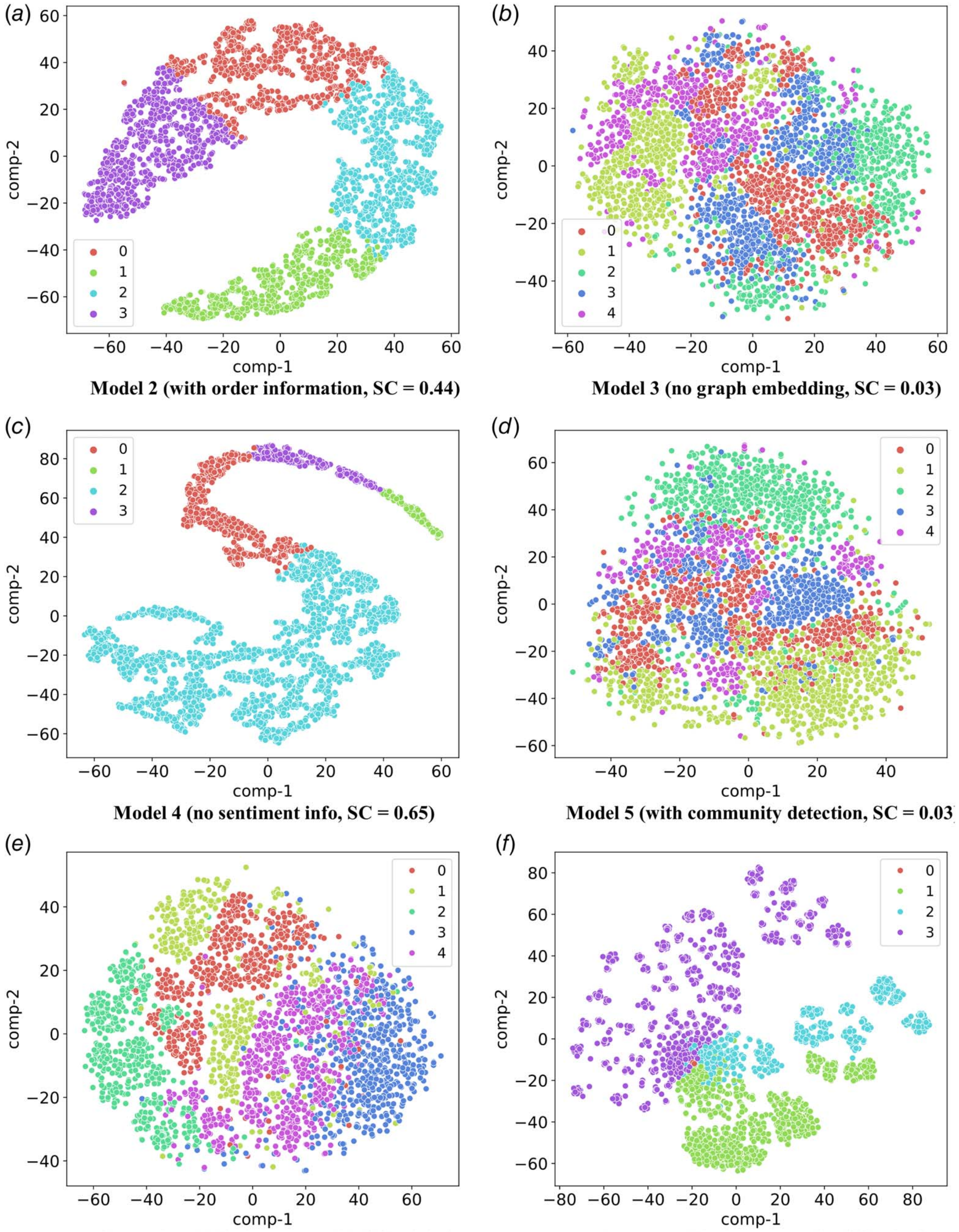
Finally, we construct an unweighted network with customer mention vectors (i.e., Model 9). This network removes the sentiment information and shows the worst performance as shown in Fig. 9(*d*). We also find that considering customer sentiments in the customer vector shows better performance than including these in the edge weights. The reason may be that customer feature vectors are directly embedded in low-dimensional customer vectors, while the edge weights are not so evident in the embeddings. Customer sentiments reflect customer characteristics which is better to be included in feature vectors.

4.3.3 Comparison of Different Types of Networks. Finally, we compare the clustering results from all tested homogeneous networks and bipartite networks. The results from the sentiment network with GAE model (Model 1) perform the best in homogeneous networks, and its performance is also better than all bipartite networks. Although the bipartite networks may contain more information, in our study, the sentiment network performs better. One possible reason is that the embedding of complex networks may be more difficult. For the bipartite networks, the information of similar customers is aggregated indirectly from the product attribute nodes, while the aggregation is direct for the sentiment network. Therefore, it may be harder to analyze the complex interrelationships between customers and product attributes, and it affects the final performance of the clustering. Another possible reason may be that since the type of the product is the same in our study, we choose product attributes as the nodes, which leads to a bipartite network of 2986 customer nodes and 28 product attribute nodes. The imbalanced number of nodes may also affect the performance of bipartite networks.

4.4 Analysis of Generated Customer Segments

4.4.1 Analysis of Customer Sentiments in Segments. After comparing different network models, the customer segments generated from the homogenous sentiment network with the GAE graph embedding model are adopted. Then, the customer needs and characteristics for each segment are analyzed and compared to find the most concerned product attributes in each customer group.

Figure 10 presents the sentiment distribution of generated customer segments, including the number of customers classified in each group and customers' average sentiment intensities to five major categories of product attributes. Here the sentiment intensities are numerical values converted from labeled sentiment polarities (i.e., "positive" to 1, "neutral" to 0, and "negative" to -1) as



Model 2 (with order information, SC = 0.44)

Model 3 (no graph embedding, SC = 0.03)

Model 4 (no sentiment info, SC = 0.65)

Model 5 (with community detection, SC = 0.03

Suryadi and Kim's method [8] (SC = 0.05)

Park and Kim's method [9] (SC = 0.10)

Fig. 7 Visualizations of customer segments generated by 6 methods: (*a*) Model 2, (*b*) Model 3, (*c*) Model 4, (*d*) Model 5, (*e*) Suryadi and Kim's method, and (*f*) Park and Kim's method. Each color represents a segment. For each method, the number of clusters with the best clustering performance on the silhouette coefficient is selected.

mentioned in Sec. 3.2.3. For each customer, the sentiment intensity to the category is the average of the sentiment intensities to the corresponding product attributes. Then, the average sentiment to categories for all customers can be calculated. We also calculated the standard errors as the error bar to examine whether there are differences in sentiments between groups.

Group 1 owns the most customers, while Group 2 has the fewest customers. At the level of product categories, most of the sentiment intensities are positive. As durable goods, vehicles are repeatedly compared before the purchase. Therefore, it is reasonable that most customers have positive feedback. Overall, there are significant differences in customer sentiments among different groups. For all categories, customers in Group 1 have the highest sentiment intensities. This group occupies the largest proportion and most customers in this group are satisfied with the vehicles they bought. One customer said, "*This car performs quite well compared to others in its class. At the same price, it offers more features; with the same features, it costs less!*"

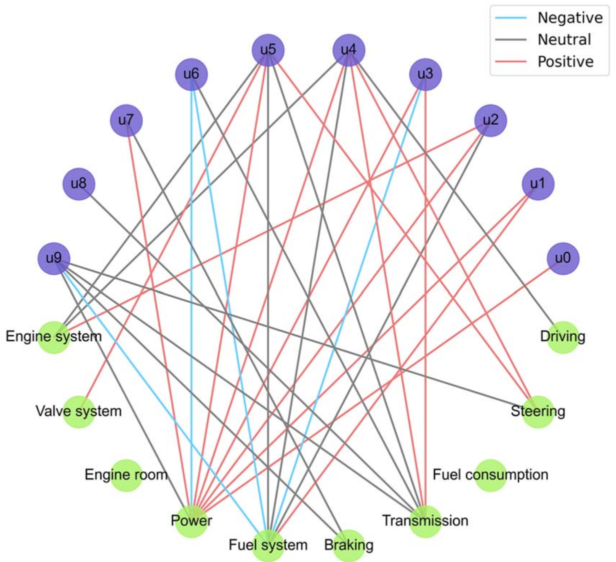
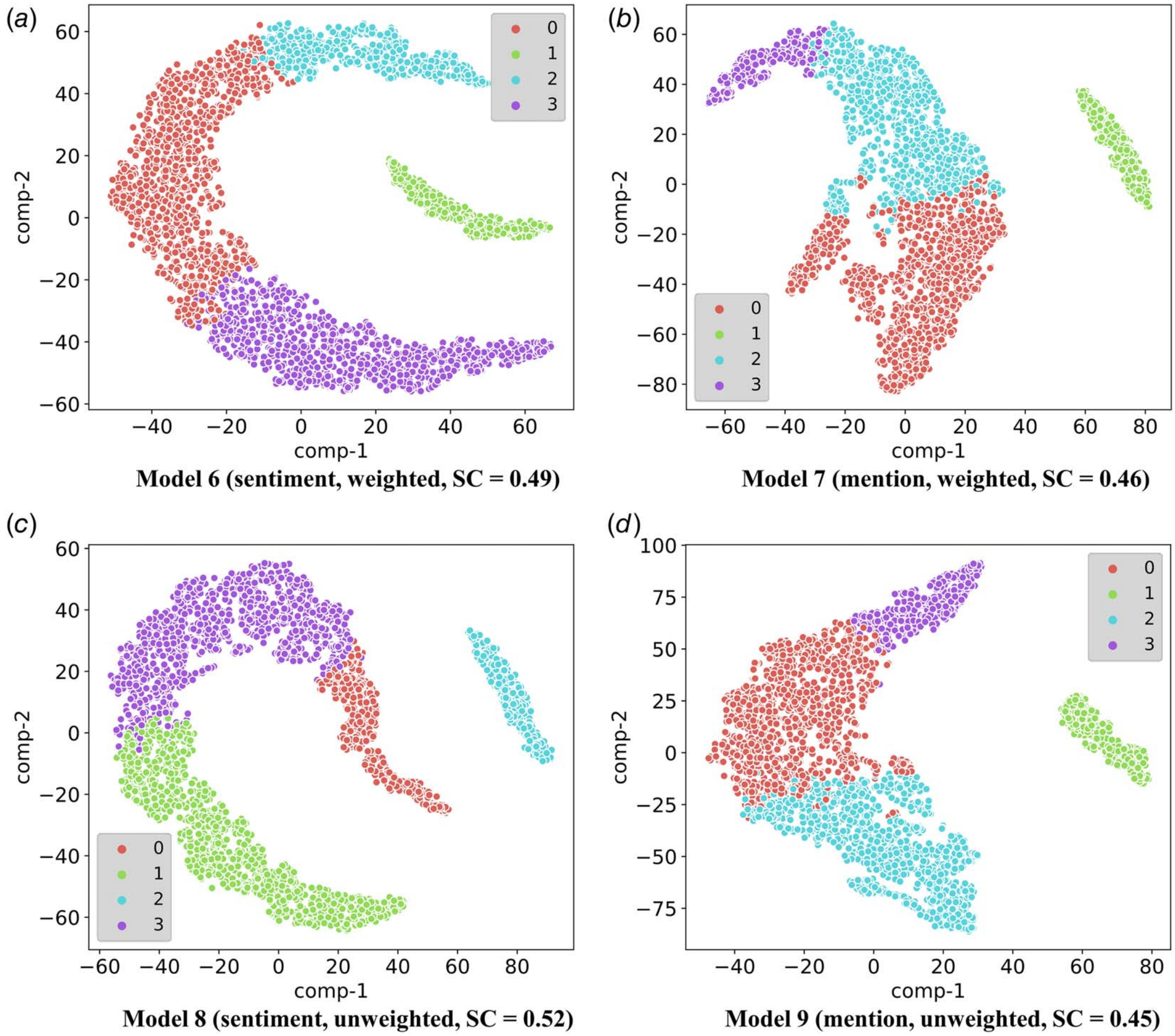


Fig. 8 Part of the constructed customer-attribute bipartite network. Purple nodes represent customers, and green nodes represent product attributes.

Customers in Group 2 have the lowest sentiment intensities in all categories, and they even have negative average sentiment in the Maintenance and comfort category. From the content of reviews from these customers, the performance of the vehicles doesn't reach their expectations, and they have more complaints. For example, a customer in Group 2 complained that "*Overtaking is a real struggle, the noise is loud, and the suspension is too stiff*." The manufacturers may focus on the group with the largest proportion and make sure most customers are satisfied with their products. Then, they can analyze the customer needs in the group with the lowest product customer satisfaction. The complaints from these customers may provide insightful suggestions for product improvement.

To analyze customer needs in more detail, we plot the average customer sentiments to the product attributes that they are concerned about. In this study, five major attributes of the vehicles (i.e., Power, Fuel consumption, Braking, Comfort, and Controllability) are selected as examples, and the plot is shown in Fig. 11. For most attributes, the average sentiments for customers in Group 1 are still the highest. The average sentiments of Power, Fuel consumption, and Controllability are positive for all customer segments, indicating the satisfaction from most customers. The average sentiments of Fuel consumption are similar among customer groups, while the sentiments are different for other attributes. Braking has the lowest performance as the average sentiment is



Model 6 (sentiment, weighted, SC = 0.49)

Model 7 (mention, weighted, SC = 0.46)

Model 8 (sentiment, unweighted, SC = 0.52)

Model 9 (mention, unweighted, SC = 0.45)

Fig. 9 Visualizations of customer segments generated by 4 bipartite network-based methods: (*a*) Model 6, (*b*) Model 7, (*c*) Model 8, and (*d*) Model 9. Each color represents a segment.

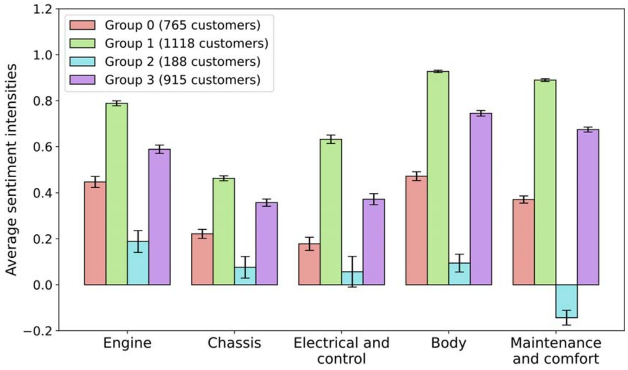


Fig. 10 The sentiment distribution of generated customer segments. The error bar represents the standard error.

neutral for customers in Group 0 and negative for customers in Group 2. If Braking is important for the design of this vehicle model, then, it may have high improvement priority.

4.4.2 Importance–Performance Analysis for Customer Segments. The importance–performance analysis is applied to each customer segment to analyze detailed customer needs. The importance and performance values for attributes are first calculated for each segment. The average customer sentiments for product attributes are calculated and normalized as the attribute performance. The importance of a product attribute is estimated by the influence of the sentiments of this attribute on customers' overall product rating from online reviews. The extreme gradient boosting (XGBoost) model is leveraged to measure this influence due to its superior performance in modeling efficiency and prediction accuracy. The input of an XGBoost model is customer sentiment scores for each attribute, and the output is the overall rating of the product. To better fit this model, customers' overall ratings on products are classified into two categories, "positive" or "negative." After the training, the importance score for each attribute is calculated by the gain-based feature importance estimation algorithm and a ten-fold cross-validation is used to obtain better modeling results. To obtain the attribute importance, four XGBoost models are built for four customer groups with prediction accuracy of 81.30%, 85.69%, 77.69%, and 91.80% in order. The final importance of product attributes is the normalized value of feature importance calculated using the XGBoost models.

Figure 12 shows the IPA plot for customers in Group 0. The vertical and horizontal blue lines are the average performance and importance for all product attributes. To compare the customer needs with the whole customer group and show the importance of customer segmentation, the average performance for all customers is calculated and plotted as the red line. From the figure, the performance of most of the product attributes is lower than the average performance for all customers. Transmission has the highest importance and its performance is below the average. With high importance but low performance, this attribute has a high priority for product improvement. Other attributes such as Multimedia, Driving, Engine system, and Fuel system should also be improved for this customer segment. Controllability, Power, and Body design perform well both in the segment and as a whole.

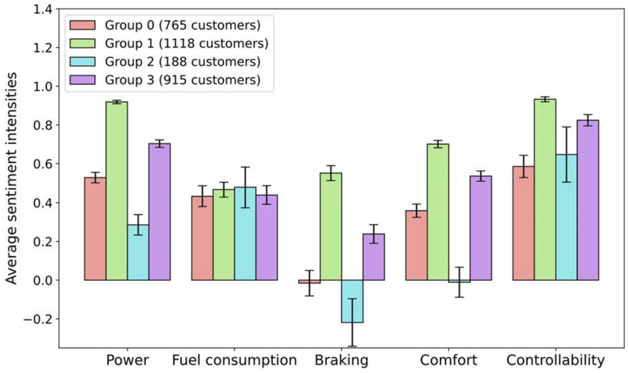


Fig. 11 The average sentiment for selected product attributes among groups. The error bar represents the standard error.

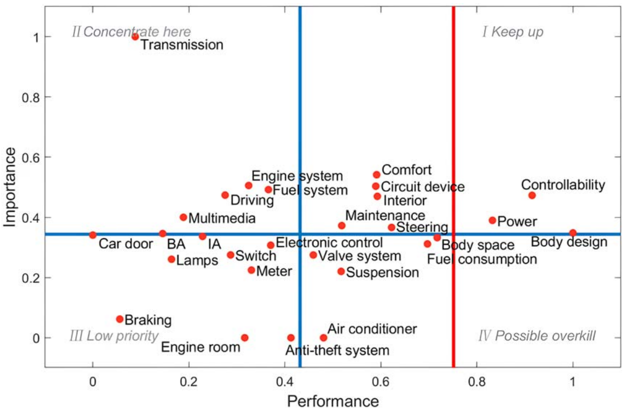


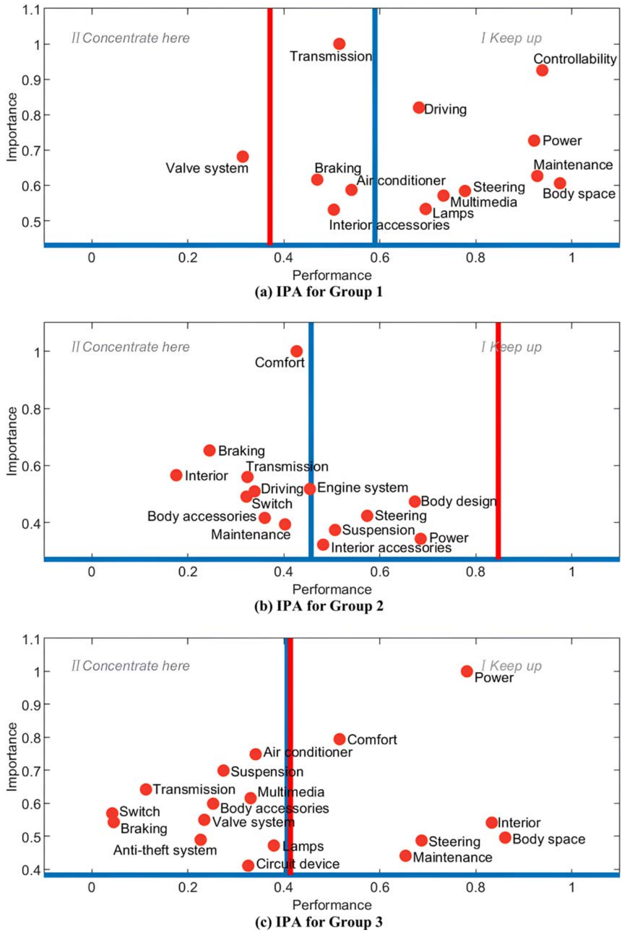
Fig. 12 The importance–performance plot for customers in Group 0. The red line is the average performance for all customers and all attributes. Here, BA stands for Body accessories and IA stands for Interior accessories.

The IPA analysis is also applied to Groups 1, 2, and 3 with similar processes. To simplify the results, we only plot the first two quadrants of the IPA plots for Groups 1 to 3 as shown in Fig. 13. For IPA plots, the attributes in the first two quadrants have an importance value greater than the average, and these attributes should be focused on for product design and improvement.

The IPA plot in Fig. 13(*a*) is for customers in Group 1. As we have mentioned, customers in Group 1 have the largest proportion and the highest average customer sentiment intensities. From the perspective of customers in Group 1, the Valve system, Transmission, Braking, Interior accessories, and Air conditioner can be improved. Similar to Group 0, Transmission still has the highest importance value. Compared with the average performance for the whole customer group, most product attributes perform well. Also, the number of attributes in the first quadrant that have high importance and high performance is the largest in all customer groups. It indicates that customers in this largest customer segment are generally satisfied with the product they bought. To improve the experience of these customers, the Valve system can be considered for improvement.

The plot in Fig. 13(*b*) is for customers in Group 2. The number of customers in Group 2 is the smallest but these customers have the lowest satisfaction with the product. From the figure, the product attributes such as Interior, Braking, Transmission, Switch, and Comfort should be improved at first. For all of the attributes with importance larger than the average, their performance is lower than the average performance for all customers. These customers tend to have negative or neutral attitudes toward the product. If the manufacturers want to control the number of negative reviews, analyzing customers in this group and focusing on the attributes with high improvement priorities may be useful.

The plot in Fig. 13(*c*) is for Group 3. The average performance of attributes is similar for this customer segment and for the whole customer group. The customers in this group are most satisfied with Power which has the largest importance and high performance. The number of attributes in the second quadrant is larger than those for other customer groups, and there are many attributes that should be improved from the perspective of these customers.



(a) IPA for Group 1

(b) IPA for Group 2

(c) IPA for Group 3

Fig. 13 The importance–performance plots for (*a*) Group 1, (*b*) Group 2, and (c) Group 3. The red line is the average sentiment for all customers and all attributes.

Since Switch, Braking, and Transmission have relatively low performance and high importance, these attributes can be improved at first.

Comparing these IPA plots, the customer needs and characteristics are quite different for customers in different customer segments. However, there are still some results in common. Power and Steering appear in the first quadrant for all customer groups. These two attributes are satisfied by most customers. Transmission and Braking have high improvement priorities for four and three groups respectively. Therefore, manufacturers should pay more attention to the improvement of these two attributes. These insights may also be useful for automakers if they expect to develop a product family with differentiated features and prices to satisfy the needs of various customer segments.

5 Conclusion

In this paper, we propose a framework for customer segmentation and need analysis based on the construction and embedding of a sentiment network of online reviewers. Customers' sentiments and purchase order information are extracted from their online reviews, and customer networks with different network architectures are constructed by examining the similarity of customer vectors. Then, low-dimensional customer vectors are embedded using Graph Autoencoder (GAE). Finally, the low-dimensional customer vectors are clustered into customer segments, and the clustering performance is compared, where the best-performed segments are fed into customer need analysis.

The methodological contribution of this work includes three-folds. First, a comprehensive framework for customer segmentation and need analysis based on the sentiment network of online reviewers and graph embedding is proposed. Our framework enables systematic processing of online reviews from product attribute extraction, customer sentiment analysis, customer segmentation, and need analysis, which can support designers in making targeted design decisions and marketing strategies. Second, we investigate the influence of using different types of information (e.g., with or without sentiment and order information) from online reviews on the segmentation performance and facilitate the clustering of high-dimensional data by leveraging graph embedding. In addition, we thoroughly examine the impact of different network structures and embedding choices on the performance of customer segmentation.

We demonstrate the proposed framework by employing online reviews of a popular sport utility vehicle in China's auto market. To evaluate the performance of our approach, nine network models are constructed using different node attribute vectors, edge weights, network types, and clustering methods. We also compared our approaches with two alternative methods from existing literature. The results indicate that the homogeneous customer sentiment network with the GAE model performs the best. Incorporating network embeddings and customer sentiment information can improve the clustering performance. After obtaining customer segments, product attribute analysis methods such as Importance–Performance Analysis are performed on each group of customers to understand their respective preferences toward product attributes. Corresponding customer choice models can be established to support the optimal selection of product attributes by maximizing the utility of a certain customer segment. This can be especially useful for product family design targeted to different segments of the market. Although we use sport utility vehicle as a demonstration in the case study, the proposed approach can be extended to the analysis of other types of vehicles or consumer electronics. We do not intend to claim our approach is the best of all, since the focus of this study is to address the challenges in existing network-based customer segmentation approaches. We hope our work can inspire researchers in design for market systems to develop more advanced methods.

One limitation of this study is that product designers have no control over which product attributes customers will comment on, and there is a risk of not covering all important product attributes solely from online review analysis. In future work, we will incorporate these factors and combine multi-source data (e.g., surveys and product maintenance records) into the segmentation model to further improve the model performance.

Acknowledgment

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (52005328) and the Shanghai Science and Technology Commission "Yangfan" Program (20YF1419300). The authors would also like to acknowledge the technical support from Yilin Zhang, Zidong Huang, and Xinyi Yang. The dataset associated with this paper is owned by the authors and has only been utilized in this study. Public can access and use this dataset only for research purposes.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The data and information that support the findings of this article are freely available.[4]

[4] See Note 3.

Nomenclature

a	=	number of product attributes
b	=	dimension of product rating
c	=	length of the order information vectors
c_0	=	cutoff point for the sentiment classification
f	=	fully connected network
j	=	index of the product attribute
k	=	number of keywords for the product attribute
p	=	center of cluster C_i
y	=	value of the element in the original adjacency matrix
P	=	final step of the iteration for bipartite GraphSAGE
A	=	adjacency matrix
E	=	set of links between nodes
F	=	dimension of customer vector
L	=	loss function
N	=	number of customer nodes
U	=	set of customer nodes
V	=	set of product attribute nodes
W	=	parameter matrices in the graph convolutional layer
X	=	feature matrix
Z	=	output of the encoder for GAE; low-dimensional node feature vector
$\hat{y}$	=	value of the corresponding element in the reconstructed adjacency matrix
$\tilde{\boldsymbol{A}}$	=	normalized Laplacian matrix
$\hat{\boldsymbol{A}}$	=	reduced graph adjacency matrix
$\tilde{\boldsymbol{D}}$	=	diagonal matrix of nodal degrees
m_i	=	point in the cluster C_i
o_{ic}	=	value of the order information for customer i
r_{ib}	=	value of the rating information for customer i
z_u	=	final embedding for customers
z_v	=	final embedding for product attributes
E_{ij}	=	link between customer i and customer j
I_N	=	identity matrix
P_j	=	product attribute vector for product attribute j
P_{jk}	=	occurrence frequency of the keyword k for attribute j
P_n	=	negative sampling distribution
Q_u	=	number of negative samples for customers
Q_v	=	number of negative samples for product attributes
S_i	=	customer vector for customer i
X_u	=	feature of customer nodes
X_v	=	feature of product attribute nodes
h_u^p	=	updated embedding of the customer node u in step p
h_v^p	=	updated embedding of the product attribute node v in step p
$h_{N(u)}^p$	=	aggregation of the neighbor nodes of the customer node in step p
$h_{N(v)}^p$	=	aggregation of the neighbor nodes of the product attribute node in step p
h_u^{p-1}	=	embedding of the customer node u in step *p-1*
h_v^{p-1}	=	embedding of the product attribute node v in step *p-1*
M_v^u	=	transformation matrices from attribute to customer
M_v^u	=	transformation matrices from customer to attribute
W_u^p	=	weight matrices of customer nodes
W_v^p	=	weight matrices of product attributes nodes
$d_a(i)$	=	average distance between point i and other points in the same cluster
$d_b(i)$	=	average distance between point i and other points in different clusters
$s(i)$	=	silhouette coefficient of a clustered point i
Neg_1	=	negative sentiment intensity generated from BiLSTM
Neg_2	=	negative sentiment intensity generated from ERNIE
$N(u)$	=	immediate neighborhood of the customer node
$N(v)$	=	immediate neighborhood of the product attribute node
Pos_1	=	positive sentiment intensity generated from BiLSTM
Pos_2	=	positive sentiment intensity generated from ERNIE
SC	=	silhouette coefficient
SSD	=	sum of squared distance
δ	=	logistic sigmoid function
γ	=	weight of negative samples
θ	=	final sentiment polarity
σ	=	sigmoid function

References

[1] Hwang, H., Jung, T., and Suh, E., 2004, "An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunication Industry," Expert Syst. Appl., **26**(2), pp. 181–188.

[2] Kim, S. Y., Jung, T. S., Suh, E. H., and Hwang, H. S., 2006, "Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A Case Study," Expert Syst. Appl., **31**(1), pp. 101–107.

[3] Wu, J., and Lin, Z., 2005, "Research on Customer Segmentation Model by Clustering," ACM Int. Conf. Proc. Ser., **113**, pp. 316–318.

[4] Toften, K., and Hammervoll, T., 2009, "Niche Firms and Marketing Strategy: An Exploratory Study of Internationally Oriented Niche Firms," Eur. J. Mark., **43**(11), pp. 1378–1391.

[5] Capon, N., Fitzsimons, G. J., and Prince, R. A., 1996, "An Individual Level Analysis of the Mutual Fund Investment Decision," J. Financ. Serv. Res., **10**(1), pp. 59–82.

[6] 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.

[7] Jiang, H., Kwong, C. K., and Yung, K. L., 2017, "A Methodology for Predicting Future Importance of Customer Needs Based on Online Customer Reviews," ASME J. Mech. Des., **139**(11), p. 111413.

[8] 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.

[9] Park, S., and Kim, H. M., 2022, "Finding Social Networks Among Online Reviewers for Customer Segmentation," ASME J. Mech. Des., **144**(12), p. 121703.

[10] Assent, I., 2012, "Clustering High Dimensional Data," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., **2**(4), pp. 340–350.

[11] Cooil, B., Aksoy, L., and Keiningham, T. L., 2008, "Approaches to Customer Segmentation," J. Relatsh. Mark., **6**(3–4), pp. 9–39.

[12] Smith, G. D., Steele, N. C., Albrecht, R. F., and Schifferl, E., 1998, "Adaptive Product Optimization and Simultaneous Customer Segmentation: A Hospitality Product Design Study with Genetic Algorithms," Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference, Norwich, UK, May. 2–5, pp. 215–218.

[13] Ertian, H., Huanhuan, L., Daqiang, C., and Yulian, F., 2013, "A Method for Customer Demands Groups Segmentation in Product Design Based on Fuzzy Clustering and Trigonometric Functions," Proceedings of the 2013 Third International Conference on Intelligent System Design and Engineering Applications, Zhangjiajie, Hunan Province, China, Jan. 16–18, pp. 95–98.

[14] Hu, X., Liu, A., Li, X., Dai, Y., and Nakao, M., 2023, "Explainable AI for Customer Segmentation in Product Development," CIRP Ann., **72**(1), pp. 89–92.

[15] Wu, R.-S., and Chou, P.-H., 2011, "Customer Segmentation of Multiple Category Data in E-Commerce Using a Soft-Clustering Approach," Electron. Commer. Res. Appl., **10**(3), pp. 331–341.

[16] Peker, S., Kocyigit, A., and Eren, P. E., 2017, "LRFMP Model for Customer Segmentation in the Grocery Retail Industry: A Case Study," Mark. Intell. Plan., **35**(4), pp. 544–559.

[17] Wang, L., Youn, B. D., Azarm, S., and Kannan, P. K., 2011, "Customer-Driven Product Design Selection Using Web Based User-Generated Content," Proceedings of the International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Washington, DC, Aug. 28–31, pp. 405–419.

[18] Jiang, S., Cai, S., Olle Olle, G., and Qin, Z., 2015, "Durable Product Review Mining for Customer Segmentation," Kybernetes, **44**(1), pp. 124–138.

[19] Joung, J., and Kim, H., 2023, "Interpretable Machine Learning-Based Approach for Customer Segmentation for New Product Development From Online Product Reviews," Int. J. Inf. Manage., **70**, p. 102641.

[20] Bondy, J. A., and Murty, U. S. R., 1976, *Graph Theory with Applications*, The Macmillan Press Ltd., London, UK.

[21] Wang, M., Chen, W., Huang, Y., Contractor, N. S., and Fu, Y., 2016, "Modeling Customer Preferences Using Multidimensional Network Analysis in Engineering Design," Des. Sci., **2**, p. e11.

[22] Bi, Y., Qiu, Y., Sha, Z., Wang, M., Fu, Y., Contractor, N., and Chen, W., 2021, "Modeling Multi-Year Customers' Considerations and Choices in China's Auto Market Using Two-Stage Bipartite Network Analysis," Networks Spat. Econ., **21**(2), pp. 365–385.

[23] Wang, H.-J., 2022, "Market Segmentation of Online Reviews: A Network Analysis Approach," Int. J. Mark. Res., **64**(5), pp. 652–671.

[24] Helal, N. A., Ismail, R. M., Badr, N. L., and Mostafa, M. G. M., 2016, "A Novel Social Network Mining Approach for Customer Segmentation and Viral Marketing," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., **6**(5), pp. 177–189.

[25] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G., 2008, "The Graph Neural Network Model," IEEE Trans. Neural Networks, **20**(1), pp. 61–80.

[26] Kipf, T. N., and Welling, M., 2016, "Semi-Supervised Classification With Graph Convolutional Networks," arXiv Prepr. arXiv1609.02907, pp. 1–3.

[27] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y., 2017, "Graph Attention Networks," arXiv Prepr. arXiv1710.10903, pp. 3–5.

[28] Hamilton, W., Ying, Z., and Leskovec, J., 2017, "Inductive Representation Learning on Large Graphs," Adv. Neural Inf. Process. Syst., **30**, pp. 1–6.

[29] Chang, S., Han, W., Tang, J., Qi, G.-J., Aggarwal, C. C., and Huang, T. S., 2015, "Heterogeneous Network Embedding via Deep Architectures," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, Aug. 10–13, pp. 119–128.

[30] Dong, Y., Chawla, N. V., and Swami, A., 2017, "Metapath2vec: Scalable Representation Learning for Heterogeneous Networks," Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada, Aug. 13–17, pp. 135–144.

[31] Gao, M., Chen, L., He, X., and Zhou, A., 2018, "Bine: Bipartite Network Embedding," Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, July 8–12, pp. 715–724.

[32] Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., and Hovy, E., 2021, "Dual Graph Convolutional Networks for Aspect-Based Sentiment Analysis," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, Aug. 1–6, pp. 6319–6329.

[33] Alamoudi, E. S., and Alghamdi, N. S., 2021, "Sentiment Classification and Aspect-Based Sentiment Analysis on Yelp Reviews Using Deep Learning and Word Embeddings," J. Decis. Syst., **30**(2–3), pp. 259–281.

[34] Xu, F., Lian, J., Han, Z., Li, Y., Xu, Y., and Xie, X., 2019, "Relation-Aware Graph Convolutional Networks for Agent-Initiated Social e-Commerce Recommendation," Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, Nov. 3–7, pp. 529–538.

[35] Salton, G., and Buckley, C., 1988, "Term-Weighting Approaches in Automatic Text Retrieval," Inf. Process. Manage., **24**(5), pp. 513–523.

[36] Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y., 2006, "A Closer Look at Skip-Gram Modelling," Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May. 22–28, pp. 1222–1225.

[37] Pelleg, D., and Moore, A. W., 2000, "X-Means: Extending k-Means With Efficient Estimation of the Number of Clusters," Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, June 29–July 2, pp. 727–734.

[38] Cheng, A., and Bi, Y., 2024, "An Integrated Data-Driven Framework for Vehicle Quality Analysis Based on Maintenance Record Mining and Bayesian Network," Int. J. Qual. Reliab. Manage.

[39] Shen, M., Cheng, A., and Bi, Y., 2024, "An Integrated Framework for Importance-Performance Analysis of Product Attributes and Validation From Online Reviews and Maintenance Records," Des. Sci., **10**, p. e19.

[40] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H., 2020, "Ernie 2.0: A Continual Pre-Training Framework for Language Understanding," Proceedings of the AAAI Conference on Artificial Intelligence, New York, Feb. 7–12, pp. 8968–8975.

[41] Xu, G., Meng, Y., Qiu, X., Yu, Z., and Wu, X., 2019, "Sentiment Analysis of Comment Texts Based on BiLSTM," IEEE Access, **7**, pp. 51522–51532.

[42] Muflikhah, L., and Baharudin, B., 2009, "Document Clustering Using Concept Space and Cosine Similarity Measurement," Proceedings of the 2009 International Conference on Computer Technology and Development, Kota Kinabalu, Malaysia, Nov. 13–15, pp. 58–62.

[43] Ye, J., 2011, "Cosine Similarity Measures for Intuitionistic Fuzzy Sets and Their Applications," Math. Comput. Model., **53**(1–2), pp. 91–97.

[44] Wang, C., Pan, S., Long, G., Zhu, X., and Jiang, J., 2017, "Mgae: Marginalized Graph Autoencoder for Graph Clustering," Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, Nov. 6–10, pp. 889–898.

[45] Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C., 2018, "Adversarially Regularized Graph Autoencoder for Graph Embedding," arXiv Prepr. arXiv1802.04407, pp. 1–3.

[46] Li, Z., Shen, X., Jiao, Y., Pan, X., Zou, P., Meng, X., Yao, C., and Bu, J., 2020, "Hierarchical Bipartite Graph Neural Networks: Towards Large-Scale e-Commerce Applications," Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, pp. 1677–1688.

[47] Likas, A., Vlassis, N., and Verbeek, J. J., 2003, "The Global K-Means Clustering Algorithm," Pattern Recognit., **36**(2), pp. 451–461.

[48] Van Der Maaten, L., 2014, "Accelerating T-SNE Using Tree-Based Algorithms," J. Mach. Learn. Res., **15**(1), pp. 3221–3245.

[49] Cheng, J.-H., Sun, D.-W., Pu, H., and Zhu, Z., 2015, "Development of Hyperspectral Imaging Coupled With Chemometric Analysis to Monitor K Value for Evaluation of Chemical Spoilage in Fish Fillets," Food Chem., **185**, pp. 245–253.

[50] Autohome, 2022, "Autohome", https://www.autohome.com.cn/

[51] Che, W., Li, Z., and Liu, T., 2010, "LTP: A Chinese Language Technology Platform," Proceedings of the 23rd International Conference on Computer Linguistics., Beijing, China, Aug. 23-27.

[52] Agnihotri, D., Verma, K., and Tripathi, P., 2014, "Pattern and Cluster Mining on Text Data," Proceedings of the 2014 4th International Conference on Communication Systems and Network Technologies. CSNT 2014, Bhopal, India, Apr. 7-9.