Big Data and Information Analytics

BDIA, 10: 1–28.
DOI: 10.3934/bdia.2026001
Received: 01 December 2025
Revised: 18 December 2025
Accepted: 19 December 2025
Published: 12 January 2026

https://www.aimspress.com/journal/BDIA

Research article

A novel comprehensive method for customer segmentation based on identifying topics and sentiments from unstructured online product reviews

Chaolong Ding and Xuesi Ma*

School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo 45400, China

* **Correspondence:** Email: maxuesi@hpu.edu.cn.

Abstract: In product development and business insights, topic extraction and sentiment analysis are crucial components. Due to the information overload in e-commerce reviews and the diverse preferences of customers, traditional research methods fail to identify commonalities among customers effectively. To overcome these challenges, we proposed an innovative five-stage ensemble approach for customer segmentation. First, TextRank was employed for data preprocessing to extract key textual features and filter relevant content. Subsequently, key topics were identified through the Word2Vec-based topic identification model. Then, to enhance the accuracy of topic-level sentiment scores, clause-level sentiment analysis was conducted using BERT, where sentiment scores were fine-tuned through TF-IDF weighting for enhanced granularity. After that, interpretable machine learning (IML) algorithms were employed to analyze user satisfaction (USAT), ensuring predictive performance and model transparency. Finally, deep embedded clustering (DEC) was leveraged to perform customer segmentation based on the extracted key topic-sentiment features. The effectiveness of the proposed method was validated through a real-world case study involving 22,320 online user reviews. The results showed that categorical boosting (CatBoost) achieved the highest performance, with an F1-score of 0.9433, demonstrating its high accuracy and transparency in predicting USAT determinants. The findings facilitate the identification of innovative product concepts.

Keywords: user satisfaction; topic identification; sentiment analysis; customer segmentation; explainable artificial intelligence

1. Introduction

In the era of big data, online product reviews are a major platform for consumers to share opinions and experiences. The rapid expansion of e-commerce has resulted in an immense volume of online product reviews. These reviews reflect customer satisfaction and unveil their perspectives and needs concerning product features. Companies can gain deeper insights into customer preferences, refine product design, optimize marketing strategies, and enhance customer satisfaction by analyzing these reviews. According to eMarketer, a leading market research firm, global e-commerce sales are projected to exceed $6.4 trillion by 2025, with 20.5% of all retail sales occurring online, highlighting the accelerating trend of digital consumption. Extracting meaningful themes and sentiment information from the massive volume of reviews has become a critical focus for researchers and industry professionals.

Traditional research on customer requirements (CRs) primarily depends on paper surveys or online questionnaires, which are often costly and time-intensive [1]. The quality of survey data heavily relies on respondents' willingness to participate and is often influenced by the length or complexity of the questionnaire. Furthermore, human-centered methods are inherently subjective, which can compromise the accuracy of the findings. Compared to traditional surveys and interviews, user-generated online reviews are highly efficient, facilitating the rapid collection of large-scale data while providing a more accurate reflection of customer interests [2]. With the rapid growth of e-commerce, the volume of customer product reviews has surged. These reviews are now a vital source of CR information, offering timely and valuable customer feedback for companies. The rise of e-commerce presents novel opportunities to enhance customer requirement analysis. Although online reviews offer advantages in terms of data volume and authenticity, their unstructured format and overwhelming information present new challenges. To overcome these challenges, we propose a systematic method for effectively extracting key information, identifying potential customer needs, and enhancing the accuracy and interpretability of analyses based on large-scale online reviews.

Traditional methods for topic identification and sentiment analysis face significant limitations, including challenges in detecting subtle topic-sentiment relationships and inefficiencies in processing large-scale data. Moreover, conventional approaches often treat topic identification and sentiment analysis as isolated tasks, limiting their ability to uncover the intricate relationships between the two in review data. Some researchers have integrated topic extraction and sentiment analysis into unified models that simultaneously capture topic and sentiment information. These topic-sentiment association models extract both types of information concurrently. A major challenge in this field is the accurate extraction of topics from unstructured, noisy text data and the capture of fine-grained sentiment information across contexts. To overcome these challenges, an ensemble method that merges topic identification with fine-grained sentiment analysis is presented in this study, offering a more precise understanding of consumer sentiments across topics. Leveraging machine learning (ML), deep learning (DL), and natural language processing (NLP) techniques, the proposed algorithm automatically identifies topics in reviews, assigns sentiment scores, and generates multidimensional feature profiles for consumers, enabling a more nuanced analysis of online reviews.

User satisfaction (USAT) is a multifaceted concept encompassing key dimensions of user experience, such as usability, perceived effectiveness, credibility, and overall value [3]. Researchers have leveraged online review data to investigate diverse domains, including household products [4], hotels [5], grocery mobile applications[3], and healthcare [6]. However, most studies depend on standard linear regression models to predict the determinants of USAT. Studies have consistently

shown that the relationship between satisfaction with product features and overall customer satisfaction is both asymmetric and nonlinear [7,8]. As the determinants are derived from topic information in online reviews, complex relationships such as multicollinearity and nonlinearity may arise between these factors and USAT. Although ML methods are effective and robust for measuring and interpreting user behavior and satisfaction from online reviews, models in USAT research often show limited predictive performance. Additionally, complex ML models face interpretability challenges due to their "black-box" nature. To address the challenge of exploring complex relationships between determinants and USAT, interpretable machine learning (IML) methods are used in this study to analyze the impact of various factors on USAT using unstructured online review data.

In this paper, we aim to propose an efficient and accurate novel methodology for customer segmentation based on identifying topics and sentiments from online product reviews. In this regard, the major contributions of this study can be summarized as follows:

(1) The Word2Vec-based topic identification model is employed in this study to improve the recognition of semantically similar words and enhance topic coherence. By leveraging the DL model BERT at the clause level and innovatively applying TF-IDF weighting, sentiment scores are calculated through a keyword-weighted averaging method under each identified topic, thereby improving the granularity and effectiveness of fine-grained sentiment scores.

(2) The interpretability of USAT predictions is improved by leveraging determinants directly extracted from the unstructured online product reviews. Using IML methods, multicollinearity is resolved while model performance is maintained, thereby advancing USAT analysis in online product reviews.

(3) Customer segmentation is enabled through Deep Embedded Clustering (DEC) analysis, with extracted key topics and their fine-grained sentiments utilized to support personalized marketing and to accurately identify target customer segments. Opportunity scores are computed based on the importance of product features and user satisfaction levels, providing data-driven insights to guide new product development and design.

(4) The large-scale Amazon Reviews dataset from the McAuley Lab is employed for metric evaluation and application demonstration. Experimental results confirm the algorithm's effectiveness and robustness, paving the way for intelligent customer management on e-commerce platforms.

The structure of this paper is as follows: In Section 2, we review relevant literature and research background. In Section 3, we detail the proposed integrated approach. In Section 4, we present a case study on an application area on Amazon. In Section 5, we discuss the theoretical and practical implications of the study and suggest future research directions. Finally, In Section 6, we conclude the study. Thus, we aim to provide novel insights and methodologies for academia and industry to leverage online reviews for customer segmentation.

2. Related work

2.1. Customer segmentation

Customer segmentation is a critical task in marketing and business analytics, designed to optimize strategies and resource allocation by identifying consumer groups with shared characteristics and needs. Basic customer segmentation methods include geographic, demographic, psychographic, and behavioral segmentation. The rise of e-commerce and social media has transformed online reviews into a vital source of consumer feedback. Assuming customer reviews represent customer opinions,

each review is characterized by the importance or emotional value of product or service features, enabling customer segmentation through Voice of Customer (VoC) vectors. Furthermore, the five-point Likert scale is applied to measure sentiment toward predefined product or service features, providing a structured description of each customer review [9]. Sentiment, feature frequency, and star ratings of product or service features are also utilized to vectorize each customer review [10]. Ultimately, VoC vectors are analyzed using K-means clustering, hierarchical clustering, or self-organizing maps to estimate customer similarities and cluster reviews to identify distinct customer segments. The primary goal of clustering algorithms is to enhance effectiveness by minimizing intra-cluster distances while maximizing inter-cluster distances. Clustering methods and outcomes are influenced by factors such as the objective function for cluster quality evaluation, assumed data structure, similarity measurement techniques, and strategies for determining the number of clusters.

Demand-based segmentation identifies diverse customer groups with varying needs, providing targeted clustering results. User sentiment clustering analysis integrates sentiment analysis with clustering to group customers effectively. Several researchers have proposed methods for customer segmentation based on preferences extracted from online reviews. Hierarchical clustering provides groupings or hierarchical structures of customer groups, but fails to represent relationships such as data correlations. To address this, Rungruang et al. [11] proposed a novel algorithm combining the RFM model and formal concept analysis (FCA). This method uncovers explicit and implicit relationships between data points, offering marketers interpretable insights into customer segments and bridging the gap between business and data science. To address the limitations of one-way clustering in generating global results and selecting segmentation variables, Wang et al. [12] introduced a bi-clustering market segmentation method that simultaneously clusters customer-related rows and pain point-related columns. This approach replaces traditional segmentation variables with customer pain points, enabling the identification of homogeneous subgroups with shared characteristics within specific subsets. Li et al. [13] combined the BiLSTM-TabNet model with the whale optimization algorithm (WOA) for customer classification, offering robust support for personalized marketing strategies. Fang et al. [14] proposed a three-stage method combining ML algorithms with RSKC clustering to systematically identify key users and their needs within complex online community networks. Guo et al. [15] introduced the LSGDM method, which leverages dual trust relationships in social networks and employs two-stage clustering. This approach builds trust networks among decision-makers through sentiment analysis and enhances group consensus using an improved Louvain algorithm, proving effective and practical for large-scale group decision-making.

Traditionally, customer segmentation mostly relied on demographic data, purchasing behavior, and survey-based questionnaires. Common approaches included rule-based grouping, traditional clustering analysis, and regression models. While these methods aid in identifying consumer groups, they exhibit significant limitations, particularly in processing unstructured data like text from online reviews. Mining information from these reviews enables the identification of consumer preferences, needs, and sentiments, facilitating more precise customer segmentation. In recent years, advancements in DL and the growing accessibility of textual data have propelled sentiment analysis from theoretical research into practical applications in NLP. This shift has significantly advanced user sentiment clustering analysis and customer segmentation. DEC algorithm is applied in this study for customer segmentation, as it simultaneously learns data representations and cluster assignments, thereby preserving information and capturing intrinsic customer relationships more effectively.

2.2. Topic extraction and sentiment analysis

Topic extraction and sentiment analysis of online reviews represent critical research areas within NLP and data mining, offering valuable applications in business intelligence and customer behavior analysis. User opinions on various topics are extracted to enable the impacts of sentiment analysis under different characteristics to be explored, with an emphasis placed on deriving actionable insights from rich online review data. The bag of words (BOW) model represents documents as vectors of word features in a high-dimensional space [16]. While BOW is simple and robust, it suffers from limitations, including a lack of semantic understanding, data sparsity, and scalability challenges. In this context, the latent Dirichlet allocation (LDA) model [17] emerged as a superior alternative, modeling documents as mixtures of topics based on content. Traditional topic models, such as LDA, are effective in identifying latent topics by modeling documents as topic distributions and topics as word distributions. However, despite its effectiveness with long texts, the LDA model struggles with data sparsity when applied to short texts like reviews or tweets. Additionally, DL has driven breakthroughs in topic extraction. Moreover, neural network-based models [18], utilizing variational autoencoder (VAE) architectures, have demonstrated superior performance in extracting topics from short texts.

Sentiment analysis of online product reviews mostly draws on advancements in NLP and statistical methodologies. This research can be categorized into three major approaches: Sentiment lexicon-based, ML-based, and DL-based sentiment analysis. Sentiment lexicon-based sentiment analysis analyzes emotions in product online reviews by aligning an emotion lexicon with the corresponding product feature lexicon. ML-based sentiment analysis utilizes labeled data and word frequency features to train classifiers that interpret product reviews, whereas DL-based sentiment analysis leverages deep neural networks to recognize complex patterns and capture contextual nuances in reviews. Jia Ke et al. [19] introduced a method integrating LDA with SnowNLP to extract feature topics and determine sentiment polarity in product reviews. Yamarthi et al. [20] proposed an entropy-based recommendation framework that integrates sentiment analysis with collaborative filtering and Bi-LSTM, where sentiment analysis of user reviews captures contextual opinions and emotions to improve personalization and address the cold-start issue in product recommendation systems. Semantic sentiment analysis focuses on identifying subjective emotional expressions, examining the relationship between sentiment polarity and product features, and applying these insights to areas such as personalized recommendations. DL models, including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformer-based architectures, have significantly advanced sentiment classification by capturing dependencies and contextual relationships among words.

Topic modeling and sentiment analysis have emerged as essential tools for user experience analysis in service and marketing domains. Jeong et al. [21] developed an opportunity-mining framework leveraging social media data through topic modeling and sentiment analysis. This approach quantified the importance and satisfaction of product themes, offering real-time tools for customer demand analysis and product planning. Xuan et al. [4] integrated the lifecycle theory with clustering analysis to extract and evaluate customer demands and green design requirements from washing machine reviews, providing theoretical and methodological insights for sustainable product design. Alsayat et al. [5] introduced a hybrid approach integrating supervised learning, text mining, and segmentation techniques, applied to TripAdvisor reviews of Mecca hotels. This method employed support vector regression (SVR), LDA, and k-means clustering to segment travelers and analyze satisfaction, offering actionable strategies for improving service quality and customer satisfaction.

Xiao et al. [22] introduced a fine-grained sentiment analysis-based user preference mining method. The method employed pre-trained language models, linguistic knowledge models, and multi-scale convolutional neural networks for user feature encoding and text representation. It framed fine-grained sentiment analysis as a sequence labeling task, demonstrating high effectiveness and practicality.

Analyzing user-generated content helps uncover consumer attitudes and needs toward products, services, and brands, supporting informed corporate decision-making. Moreover, researchers have developed advanced algorithms to accurately identify latent topics in reviews and capture user sentiment tendencies. However, the diversity of expressions and the complexity of emotional information in review data pose significant challenges to topic extraction and sentiment analysis. Topic modeling identifies hidden thematic structures in extensive online reviews, while sentiment analysis uses polarity-detection algorithms to quantify emotions. Specifically, Word2vec embeds words into a low-dimensional vector space using two primary techniques: skip-gram and continuous bag of words (CBOW). The Skip-gram model predicts context words given a target word, whereas CBOW predicts a target word based on its surrounding context. Word2vec's efficient log-linear neural network, compact vector representation, resource-friendly implementation, and versatility in downstream NLP tasks have made it widely adopted. BERT embeddings effectively handle synonyms and polysemous words by leveraging contextual word meanings. A key technical feature of BERT is its use of bidirectional training, which is derived from the Transformer attention mechanism widely adopted in machine translation. In this study, we employ the BERT-base-multilingual-uncased-sentiment model to analyze sentiments associated with the identified determinants of USAT.

2.3. The impact of determinants on USAT

In recent years, driven by digital transformation and the growing use of intelligent technologies, USAT has gained significant attention in fields such as e-commerce, services, and product design. Researchers have identified several key determinants of user satisfaction, such as ease of use, perceived usefulness, security, and emotional satisfaction. These factors influence user satisfaction to varying extents across domains and applications. While recent ML models have shown high predictive performance in various data analyses, they often lack interpretability in their predictions. Studies on USAT in reviews have also confirmed this trend. Therefore, interpretability is needed to enable users to translate model results into understandable terms.

Research on USAT helps companies optimize user experience and provides strong support for market segmentation, product design, and customer service. Researchers have explored factors influencing USAT across fields using online user data. Wang et al. [23] proposed a method to generate game experience dimensions by processing online reviews of video games automatically. They applied the LDA algorithm to extract themes from the review texts and used regression analysis to examine the relationship between game components and user satisfaction. Liu et al. [24] introduced a customer requirement (CR) modeling framework based on the BW-CNN and S-Kano models, analyzing customer needs in mobile game reviews and their impact on satisfaction. This framework provides game developers with suggestions for improvement and validates its effectiveness and heterogeneity in identifying customer needs across games. Park [25] introduced a new method that combines data envelopment analysis (DEA) with text mining to assess the level of customer satisfaction (LCS). The DEA model was applied to assess the LCS of service providers. Kumar et al. [26] used LDA, regression analysis, and explicit analysis to identify ten determinants

of user satisfaction for the Amazon Alexa application, offering valuable business insights for enhancing the competitive advantage of voice assistants.

Despite identifying several key factors influencing USAT, studies still face challenges. As technology advances, user demands and expectations evolve, complicating satisfaction determinants. This requires more precise data collection and processing techniques, especially in multi-dimensional sentiment analysis and user behavior modeling. Studies have shown that the relationship between satisfaction with product features and overall customer satisfaction is asymmetric and nonlinear [27]. While ML can capture these nonlinear relationships, its "black box" nature complicates understanding how satisfaction with specific product features affects overall customer satisfaction. Therefore, we propose a method that combines ML with the explainable artificial intelligence SHapley Additive exPlanations (SHAP) method to explain the nonlinear relationship between product feature sentiment in each review and overall customer satisfaction, thus bridging the gap between high predictive capability and interpretability. Thus, in this study, we present ten ML models to predict the factors influencing USAT in reviews and select algorithms with high predictive performance through comparisons. Additionally, by applying the SHAP method to the constructed models, the key features contributing to USAT in reviews are identified.

3. Methodology

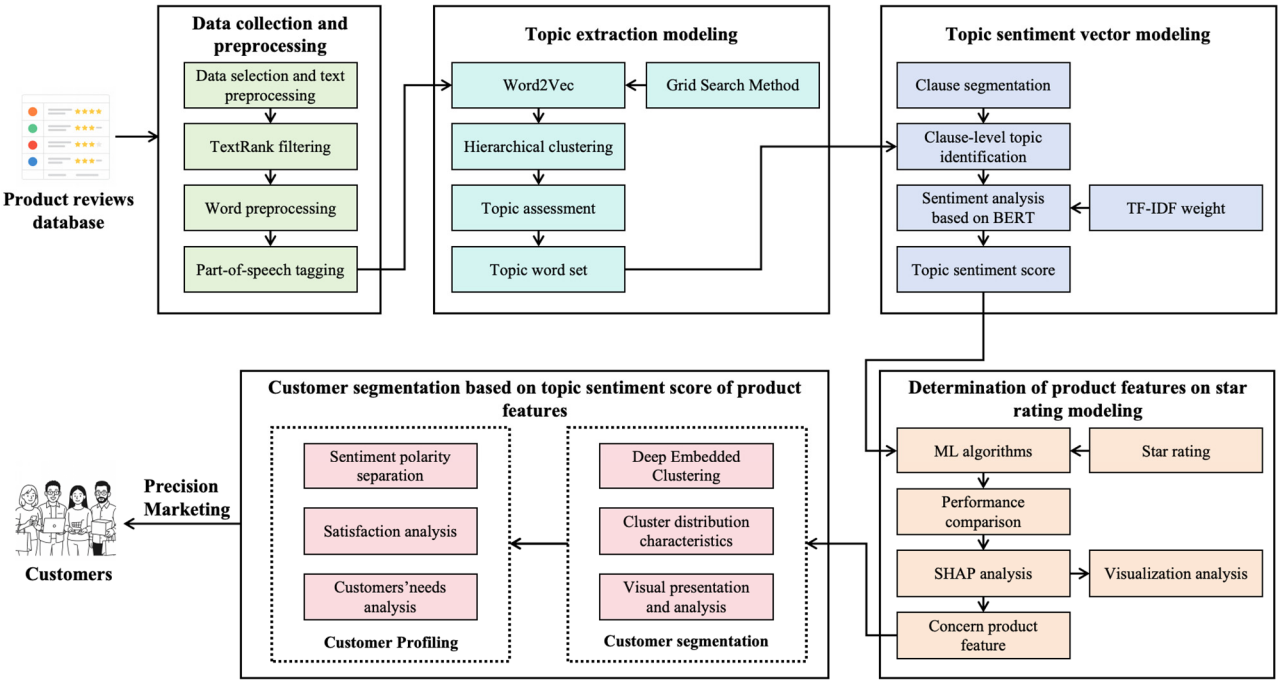

Figure 1. Overall workflow of the proposed integrated system.

To enhance the accuracy of topic-level sentiment scores, improve the interpretability of ML algorithms in USAT prediction, and address the limitations of fine-grained customer segmentation on large-scale unstructured online reviews data, we introduce a comprehensive method integrating TextRank, Word2Vec-based topic model, TF-IDF, BERT-based sentiment analysis model, IML

algorithms, and Deep Embedded Clustering. The overall workflow of the proposed method is depicted in Figure 1. To achieve this goal, we analyze online product reviews. The use of online textual reviews offers the advantage of abundant, unrestricted commentary, unbound by the constraints of surveys or standardized processes. Amazon, one of the world's leading e-commerce platforms, is the data source for this study, utilizing its product reviews. The proposed method consists of five sequential stages: (1) Data collection and preprocessing, (2) Topic extraction modeling, (3) Topic sentiment vector modeling, (4) Determination of product features on star rating modeling, and (5) Customer segmentation based on topic sentiment score of product features.

3.1. Data collection and preprocessing

For data collection, unstructured online product reviews are obtained from e-commerce websites, particularly Amazon, the world's leading online shopping platform. Each review contains details including the reviewer's user ID, review title, review text, star rating, review date, helpfulness score, and verified purchase status. The dataset contains five-star ratings ranging from 1 (very dissatisfied) to 5 (very satisfied), serving as an indicator of user satisfaction. To ensure representative customer feedback, the proposed method emphasizes the inclusion of a substantial number of verified purchase reviews.

Before topic extraction, the TextRank algorithm is applied to extract keywords from review text, filtering relevant content. TextRank, adapted from the PageRank algorithm, modifies directed edges into undirected edges [28]. TextRank algorithm breaks text into minimal units—words—and treats each word as a "node on the World Wide Web". It calculates the importance of each word based on its co-occurrence relationships with other words. The core formula of the TextRank algorithm is presented below:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in in(V_i)} \frac{W_{ji}}{V_{k \in out(V_j)} W_{ik}} WS(V_j), \tag{1}$$

where, W_{ji} denotes the weighted edge between two nodes, indicating their varying degrees of importance. The damping factor d controls the probability of a random jump between nodes. The set $in(V_i)$ consists of nodes pointing to V_i while $out(V_j)$ represents nodes to which V_j points. W_{ji} indicates the weight of the edge directed from V_j to V_i.

Preprocessing the dataset is a critical step in converting text into a vector space representation, which enables statistical analysis. Initially, textual reviews are cleaned by removing poorly formatted entries, missing content, punctuation, and extraneous characters, including the "@" symbol and URLs linking to images and videos. Subsequently, all text is converted to lowercase. Tokenization is then applied to segment the cleaned reviews into individual words. Words such as stop words, single-character words, words containing numbers, those with excessive repetitive characters or underscores, and misspelled words are removed. Next, part-of-speech (POS) tagging [29] is used to organize the reviews into preprocessed word classes, including nouns, verbs, adjectives, and adverbs, facilitating the identification of product features. Given that previous studies [30] often consider nouns as product feature words, this study selects nouns as candidate feature words.

3.2. Topic extraction modeling

The word embedding-based approach is employed for topic extraction, enabling the automatic identification and grouping of product feature terms based on specified hyperparameters of word

embedding vectors. Word2Vec is utilized as the word embedding method to vectorize nouns extracted from reviews [31]. Word2Vec encodes words from reviews as vectors derived from their context (i.e., surrounding words). In the Word2Vec model, word dimensions, window size, and truncation frequency are configured as hyperparameters. Hierarchical clustering is subsequently applied to group similar nouns based on their corresponding vectors. These nouns serve as representative terms for the clustered topics.

Refinement is conducted to enhance the coherence and separation of clustered topics. Clustered topic coherence is measured by the cosine similarity between nouns within a cluster, indicating the internal consistency of words representing the topic. The coherence score for each topic is calculated by evaluating the similarity among the top 20 keywords to assess its validity. Clustered topic separation is defined as the distance between clusters and is calculated using the average cosine similarity among the top 20 keywords of each topic. If the similarity between representative keywords of different topics surpasses a predefined threshold, the clusters are merged to improve topic coherence. For instance, if the similarity between "Topic 1" and "Topic 5" exceeds the threshold 0.9, the two topics are combined into one. This refinement process enhances within-cluster similarity among nouns and reduces redundancy across topics.

3.3. Topic sentiment vector modeling

Factors contributing to satisfaction may differ from those driving dissatisfaction [32]. Here, we conduct sentiment analysis to compare identified product feature topics and evaluate their effects on satisfaction and dissatisfaction. The pre-trained BERT-base-uncased-sentiment model is employed to compute fine-grained sentiment scores for these product feature topics and their corresponding review texts. Specifically, the nlptown/bert-base-multilingual-uncased-sentiment model, hosted in the official Hugging Face repository (https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment), is utilized. The model has been fine-tuned for sentiment analysis of product reviews in six languages: English, Dutch, German, French, Spanish, and Italian. It predicts sentiment as star ratings.

TF-IDF is a widely used technique in information retrieval and text mining. It assesses the importance of a word within a document relative to its occurrence across a collection of documents [33]. The fundamental concept of the TF-IDF algorithm is as follows:

$$TF_i = \frac{n_{ij}}{\sum_k n_{kj}}, \tag{2}$$

$$IDF_i = log\frac{|D|}{|D_i|+1}, \tag{3}$$

$$TF - IDF_i = TF_i \times IDF_i. \tag{4}$$

Term frequency (TF): TF represents the frequency of a specific word appearing in a document, where n_{ij} denotes the frequency of the j_{th} word in the i_{th} document.

Inverse document frequency (IDF): IDF quantifies the significance of a word relative to its distribution across the document collection. Here, D indicates the total number of documents in the collection, while D_i represents the number of documents containing the word i.

TF-IDF: TF-IDF integrates term frequency and inverse document frequency into a single score, reflecting the significance of a word in a specific document. The TF-IDF value is positively correlated

with the term frequency in the text and negatively correlated with its frequency in the corpus. This implies that a term with high term frequency and low inverse document frequency will yield a high TF-IDF value. A higher TF-IDF value indicates stronger discriminative power of the term and greater representativeness within the text.

In this study, review sentiment is not limited to classifications of positive, negative, or neutral; instead, it is quantified as a continuous value reflecting degrees of positivity or negativity. Sentiment scores are represented as continuous values ranging from 0 (negative) to 1 (positive), enabling a more granular analysis of their influence on review helpfulness. By applying a fine-tuned BERT model at the clause level and weighting sentiment scores using TF-IDF based on the keywords extracted under each topic, the granularity of sentiment scoring is significantly enhanced. The proposed topic sentiment vector modeling method comprises four steps, each corresponding to a process, as outlined below:

(1) Candidate feature word assignment. The similarity between word vectors and each topic's central vector is calculated to assign words to their most relevant feature topic. Each candidate feature word is assigned to its most relevant feature topic. Initially, topic words are used for the assignment. Subsequently, an optimization process refines the allocation by comparing the similarity between word vectors and topic vectors, ensuring each word is assigned to the most appropriate topic.

(2) Construction of topic word weights. First, the TF-IDF values of all keywords in the review text are calculated. Next, for each feature theme, the total TF-IDF value is obtained by summing the TF-IDF scores of all keywords associated with that theme. Finally, the TF-IDF weight of each word is normalized as a proportion of the total TF-IDF value of its feature theme. This normalization provides the relative importance of each word, represented as a word weight vector $w = \{w_1, w_2, \cdots, w_m\}$ for each feature theme.

(3) Clause extraction and construction. Research has demonstrated that performing sentiment analysis at the clause level for short texts, such as e-commerce reviews, can overcome the limitations of traditional sentence-level models, which often fail to capture the multiple factors present in a single review [19]. For feature mining and sentiment analysis, let $R = \{r_1, r_2, \cdots, r_n\}$ denote the collection of reviews for a product. Initially, each preprocessed review is segmented into sentences, which are then further split by commas, semicolons, and colons to generate an initial set of clauses $S = \{s_1, s_2, \cdots, s_n\}$. Feature words within each clause are identified to form a feature word set, $T_i = \{t_1, t_2, \cdots, t_k\}$, for the clauses in S_i.

(4) Clause-based fine-grained sentiment value calculation. We present an optimized method for calculating fine-grained sentiment scores. Initially, we use the BERT model to extract both the overall sentiment score of the review text and the sentiment scores for individual clauses. Keywords for each clause are identified, and the weighted sentiment scores are computed using TF-IDF word weights. We then compute the keyword-weighted average of the clause sentiment scores for each theme. Finally, we derive the fine-grained sentiment values $E_i = \{e_1, e_2, \cdots, e_n\}$ for all themes for each user. This method improves the accuracy of sentiment scores at the theme level by refining the sentiment analysis process.

3.4. Determinants' influence on USAT with star ratings

In line with other research [34], we use the 5-star ratings provided by users as ground truth labels to evaluate the performance of the sentiment analysis model. The star ratings are converted into two labels (i.e., negative and positive), as the classifier's predictive capability is greater when predicting two labels compared to five-star ratings [30]. One-star and two-star ratings are labeled as negative,

while four-star and five-star ratings are labeled as positive. Three-star ratings are classified as either positive or negative based on the sentiment expressed in the review.

Ten ML algorithms are evaluated, including extreme gradient boosting (XGBoost) [35], random forest (RF) [36], adaptive boosting (AdaBoost) [37], gradient boosting decision tree (GBDT) [38], bootstrap aggregating (bagging) [39], voting regressor [40], logistic regression (LR) [41], light gradient boosting machine (LGBM) [42], categorical boosting (CatBoost) [43], and support vector machine (SVM) [44], to identify the optimal classifier. These classifiers are designed to predict user satisfaction based on sentiment scores derived from product features. In constructing a ML classifier, the input variables consist of fine-grained sentiment scores for product feature topics in online reviews (e.g., sentiment scores $\mathrm{pfs}_1, \mathrm{pfs}_2, \cdots, \mathrm{pfs}_I$ corresponding to each user ID), The output variable is the review's star rating label, classified as either positive or negative (Table 1). Accuracy, precision, recall, and F1-score are used to comprehensively evaluate the performance of our sentiment analysis model.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}, \tag{5}$$

$$Precsion = \frac{TP}{(TP+FP)}, \tag{6}$$

$$Recall = \frac{TP}{TP+FN}, \tag{7}$$

$$F1-Score = 2 \times \frac{Precsion \times Recall}{Precsion+Recall}. \tag{8}$$

Accuracy, defined as the proportion of correct predictions, is regarded as the most reliable metric for evaluating algorithms. TP represents the number of samples correctly predicted as positive, TN represents the number of samples correctly predicted as negative, FP refers to false positives, where the model incorrectly predicts positive, and FN refers to false negatives, where the model incorrectly predicts negative. Precision is defined as the proportion of true positives correctly identified as positive. Recall, also known as sensitivity or true positive rate (TPR), represents the proportion of actual positive cases correctly identified by the model. The F1-score combines Precision and Recall, defined as their harmonic mean.

Table 1. Fine-grained sentiment scores for all topics of each user.

User id	Rating	pfs_1	pfs_2	...	pfs_I	Label
AGS7ZVB4DBRNSCE6RJ65BLSNIKAA	5	0.7900	0.6966	...	0.7900	Positive
AF667PU25AUN6EGTCNPJVMQYIVSA	4	0.7167	0.3137	...	0.6637	Positive
AFP6ADTSWTWV42SASXQ5RHCIREHQ	3	0.4850	0.5983	...	0.4583	Positive
AGYXTURXYAE7P4SOTV2ESWGBYVHQ	3	0.4631	0.5665		0.3963	Negative
AGIKMYZJQOK5ADXHWA7JXZLE3QBA	2	0.2356	0.5296	...	0.2590	Negative
AEEG3LWNUCTZPKSS6HV5FDI6SIEA	1	0.1937	0.0852	...	0.1402	Negative
...						

Interpretable machine learning (IML) offers clear explanations by interpreting black-box models. IML can be categorized into two types based on its characteristics: intrinsic interpretable models and post hoc explanation methods [45]. Intrinsic interpretable models, like linear models and decision trees, are interpretable because of their simple structures. Post hoc explanation methods, however, focus on techniques for constructing explanations after the model is built. Compared to intrinsic models, post

hoc methods can freely use various ML models, as the model and its explanation are independent. Here, we apply the post-hoc explanation technique SHAP to enhance model interpretability by accounting for interactions among all possible features. This approach enables a comparative understanding of how different feature topics influence users' overall satisfaction.

SHAP is a method for explaining individual predictions using optimal Shapley values from game theory. SHAP values measure each feature's contribution to the model [46]. As shown in the formula, the Shapley value (ϕ_i) for the $i-th$ data point is calculated by considering all possible subsets ($S \subseteq F$)) of features (F) in the learning process. The value $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)$ is computed by subtracting the prediction of the contribution model that excludes feature i from the prediction of the model that includes feature i. The Shapley value is averaged over all possible feature combinations, excluding those that omit feature x_S. The SHAP model decomposes and quantifies the contribution of each feature [47].

$$\phi_i = \sum_{S\subseteq\frac{F}{\{i\}}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left(f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right). \tag{9}$$

We combine ML algorithms with SHAP-based interpretable models to quantify the importance of each feature theme in shaping overall user satisfaction. The technique interprets prediction results by assigning Shapley values to feature themes according to their specific attributes. Effective interpretation requires ensuring the explanatory model's local accuracy, mitigating the impact of missing features in the input, and preserving feature importance for highly relied-upon features without interference from others [48].

3.5. Customer segmentation based on a topic sentiment score of product features

3.5.1. Deep embedding clustering algorithm

Deep embedding clustering (DEC) is an algorithm integrating DL, an autoencoder (AE), and the K-means clustering technique [49]. The algorithm utilizes the encoder layer of the AE to produce embedded data representations, which are subsequently clustered using K-means within the embedded space. Furthermore, DEC incorporates the t-distribution from t-SNE to convert clustering metrics into probability values, with the metrics calculated as follows:

$$q_{ij} = \frac{\left(1+\frac{\left\|z_i-\mu_j\right\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1+\frac{\left\|z_i-\mu_{j'}\right\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}, \tag{10}$$

where z_i represents the embedded representation of the i_{th} sample generated by the encoder, μ_j denotes the center of the j_{th} cluster, and α is the degrees of freedom parameter in the Student's t-distribution. q_{ij} denotes the probability that sample i belongs to cluster j.

The DEC algorithm introduces an auxiliary distribution, using the KL divergence to quantify the difference between this auxiliary distribution and the actual distribution, which forms the basis of the loss function. p_{ij} denotes the probability that the adjusted data point i is associated with the cluster

center j. The auxiliary distribution and the corresponding loss function L are calculated as follows:

$$p_{ij} = \frac{\frac{q_{ij}^2}{\sum_i q_{ij}}}{\sum_j \left(\frac{q_{ij}^2}{\sum_i q_{ij}}\right)}, \tag{11}$$

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}. \tag{12}$$

The DEC algorithm comprises two stages: First, a pre-trained AE optimizes the model by minimizing reconstruction loss. Second, the encoder of the AE is linked to the clustering layer for clustering training. During clustering training, the encoder is refined by minimizing clustering loss and reconstruction loss, which yields the clustering results. The structural diagram of the DEC algorithm is presented in Figure 2.

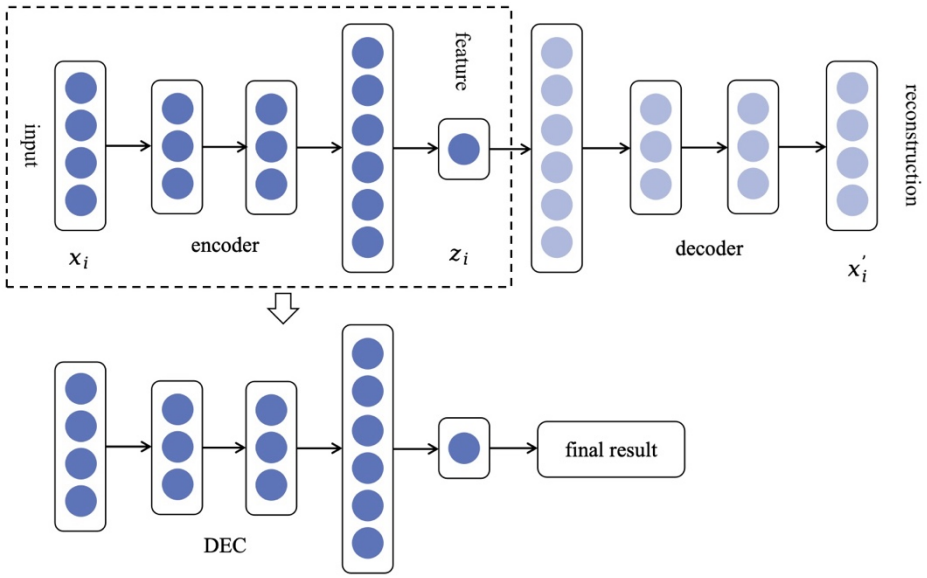

Figure 2. The structure diagram of the deep embedding clustering algorithm.

Compared to traditional clustering algorithms, the DEC algorithm leverages an AE to preserve high-dimensional data characteristics more effectively. This approach reduces interference from irrelevant features and significantly accelerates clustering processes [50]. By employing a center-of-mass-based probability distribution and adopting KL divergence as the clustering loss function, DEC demonstrates superior clustering performance in practical applications.

3.5.2. Customers' needs analysis

In the proposed method, the IML method SHAP is applied to identify key determinants of USAT based on feature importance, which are treated as product features for customer segmentation. The DEC algorithm then segments customer reviews based on the fine-grained sentiment scores of these product features. Using the importance and satisfaction data of specific product features in these segments, companies can develop new products tailored to each market segment by optimizing features favored by different customer groups. For each product feature i in each customer group c,

an opportunity score is calculated using its associated importance and satisfaction levels, forming a foundation for new product development [51]:

$$Opportunity_{ci} = Importance_{ci} + Max(Importance_{ci} - Satisfaction_{ci}, 0), \quad (13)$$

where $Importance_{ci}$ represents the average importance of each product feature within cluster c, derived from SHAP values. $Satisfaction_{ci}$ denotes the average sentiment value of each product feature within cluster c. To reduce the influence of extreme values on the resulting opportunity scores, the importance and satisfaction scores for product feature i in each customer group c are first scaled using quantile-based scaling and subsequently rescaled to the 0–10 range [21]. Consequently, the opportunity score ranges between 0 and 20. Product features with an opportunity score of 10 or above are identified as potential market opportunities. These features are crucial for overall customer satisfaction but fall short of meeting the needs of the customer group. Enhancing these features in the next generation of products can address unmet needs, attracting potential customers. This opportunity score enables companies to quantitatively identify new product development opportunities within customer segments.

4. Case study

4.1. Data acquisition and preprocessing

We leverage the Musical Instruments category from the large-scale Amazon Reviews dataset, published by McAuley Lab on the Hugging Face repository, to conduct sentiment analysis on e-commerce reviews. Only English comments with more than one helpful vote are included to ensure assessment reliability. To ensure each review reflects a unique customer opinion, duplicates are removed by checking for identical entries in the author, title, and content fields. A total of 24,898 reviews, published between January 2022 and August 2023, are collected.

A preliminary assessment is performed to evaluate the relevance of selected texts to the domain. Filtering with TextRank reduces the total number of reviews to 22,320, all of which contain the extracted keywords. After data processing, exploratory data analysis is conducted. A word cloud, illustrating the distribution of product feature candidate words, is generated based on word frequency in the preprocessed review data, as shown in Figure 3.

4.2. Identification and extraction of USAT determinant

The hyperparameters of Word2vec dimension (300), window size (3), and word cutoff frequency (5), are selected based on the lowest Davies-Bouldin index (DBI), achieving optimal clustering results (Table 2) [52]. The similarity threshold is applied to refine clustering coherence and separation. The initial topic count is set to 10, and coherence scores are calculated by evaluating the similarity among the top 20 keywords within each topic. A higher coherence score indicates more meaningful topics, with the average cosine similarity among the top 20 keywords per cluster consistently exceeding 0.9. Topics are merged using a separation threshold, decreasing the average cosine similarity between clusters from 0.5208 to 0.4796. Finally, nine product attributes are identified from the online reviews (

Table **3**).

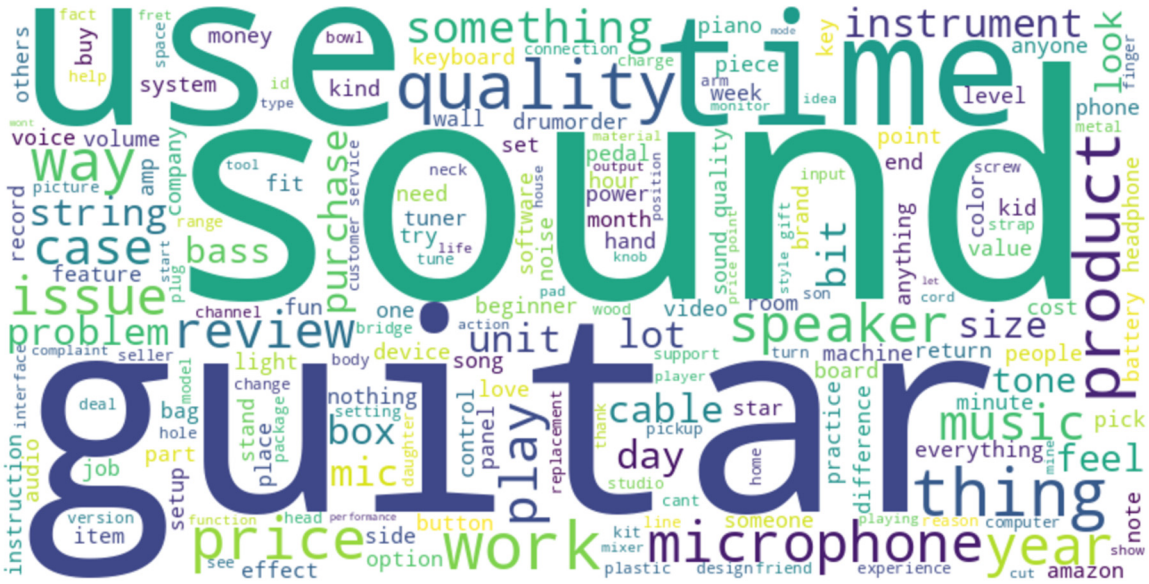

Figure 3. Word cloud image after preprocessing Musical Instruments review data.

Table 2. Clustering performance based on combining Word2vec hyperparameters.

Dimensions	Window size	Cutoff	DBI index
100	3	5	1.6558
100	3	10	1.7904
100	4	5	1.5762
100	4	10	1.6603
100	5	5	1.5860
100	5	10	1.6066
200	3	5	1.5148
200	3	10	1.5798
200	4	5	1.5985
200	4	10	1.7509
200	5	5	1.5409
200	5	10	1.7030
300	**3**	**5**	**1.4768**
300	3	10	1.4936
300	4	5	1.5049
300	4	10	1.7010
300	5	5	1.6757
300	5	10	1.5932

Table 3. The top 20 words under each topic.

Topic	Determinant	Keywords per topic
Topic1	Audio equipment and performances	singers, broadcast, performances, downside, meeting, sync, tablets, capture, radios, booth, ceremony, clear, podium, plugin, enhancements, media, flash, mixing, earpiece, crackles
Topic2	Mechanical components and adjustments	thick, thicker, thinner, top, spring, mechanism, lock, nuts, slide, loosen, tightening, sand, bolt, bearing, markers
Topic3	Instruments and music education	adult, toy, child, kid, interest, grandson, boy, learning, learn, granddaughter, student, age, flute, lessons, trumpet, saxophone, husband, children, loves, plethora
Topic4	After-sales service and warranty	warranties, returns, fee, saw, credit, fault, hesitation, replacements, communication, give, ship, account, happens, vendor, address, retailer, owner, warehouse, companies, bazaar
Topic5	Audio equipment connections and interfaces	devices, source, inputs, mixer, ports, jacks, connection, port, input, dongles, zoom, stereo, phantom, outputs, connects, interface, channels, mono, latency, computer
Topic6	String instrument parts	intonation, truss, rod, fret, ends, frets, string, saddles, bone, buzz, nut, headstock, tuners, tension, neck, carbon, fiber, saddle, adjustment, stock
Topic7	Audio effects processing and adjustment	boost, unity, distortion, chorus, delay, presets, controls, effect, master, volumes, drive, crunchy, reverb, levels, compression, amps, modes, passive, level, crank
Topic8	Customer service and shopping	company, seller, amazon, refund, warranty, replacement, return, email, contact, manufacturer, review, rating, star, questions, process, order, hope, window request, support
Topic9	Performances and experiences	shine, death, dislike, workers, upwards, scene, locations, rattle, term, bands, horns, staple, fragile, rocket, city, kept, ram, sabers, concept, aspects

4.3. Evaluation of ML algorithms

The summary of the performance metrics for algorithms predicting overall satisfaction is provided in Table 4. CatBoost achieved the highest F1-score (0.9433), followed by XGBoost (0.9427), GBDT (0.9424), Bagging (0.9417), LGBM (0.9410), Adaboost (0.9406), RF (0.9406), VotingRegressor (0.9200), and SVM (0.9219). In contrast, LR had the lowest F1-score (0.9143), indicating relatively weaker performance than the other algorithms. Nonetheless, all algorithms showed strong predictive capabilities, emphasizing the robustness of our sentiment analysis model. CatBoost also exhibited the highest accuracy (0.9138) in predicting overall satisfaction, outperforming the other algorithms in predictive performance. XGBoost (0.9130) and GBDT (0.9125) closely followed in accuracy. The superior accuracy of CatBoost highlights its suitability for this dataset compared to other models. In summary, CatBoost emerged as the most suitable model for this dataset. Consequently, the IML interpretation with CatBoost was deemed the most accurate for this dataset.

Table 4. Comparison of prediction performance based on different machine learning algorithms.

Model	Accuracy	Precision	Recall	F1-score
LR	0.8875	0.9023	0.9264	0.9143
SVM	0.8925	0.9148	0.9294	0.9219
VotingRegressor	0.8775	0.9195	0.9206	0.9200
RF	0.9093	0.9438	0.9374	0.9406
Adaboost	0.9092	0.9414	**0.9399**	0.9406
LGBM	0.9105	0.9498	0.9323	0.9410
Bagging	0.9116	0.9500	0.9336	0.9417
GBDT	0.9125	0.9497	0.9352	0.9424
XGBoost	0.9130	**0.9519**	0.9336	0.9427
CatBoost	**0.9138**	0.9499	0.9368	**0.9433**

4.4. Effects of product features

Among all evaluated machine learning algorithms, CatBoost, XGBoost, and GBDT achieved the highest F1-scores and were therefore selected for SHAP-based interpretability analysis. The average influence of each feature on the model's output magnitude is shown in Figure 4, with features ranked according to their level of impact. Across all three models, the most critical factors for predicting USAT consistently included customer service and shopping (CSS), string instrument parts (SIP), audio equipment connections and interfaces (AECI), audio effects processing and adjustment (AEPA), and performances and experiences (PE), instruments and music education (IME), after-sales service and warranty (ASSW). To enhance the reliability and robustness of these key features, feature importance is evaluated across well-performing models, and key determinants are identified using an intersection strategy. These product feature topics are identified as key determinants necessitating further investigation into overall user satisfaction. Detailed analysis revealed the interactions between these feature topics and overall user satisfaction.

Figure 5 illustrates the SHAP value distribution for the six most impactful product feature topics on user satisfaction: CSS, SIP, AECI, AEPA, PE, and IME. This analysis provides insights into how these topics influence user satisfaction mechanisms. The horizontal axis denotes feature topic values, while the vertical axis shows their corresponding SHAP values. This visualization reveals how the contribution of each feature topic to the outcome changes with variations in feature topic values.

The SHAP values for SIP, AECI, and AEPA exhibited a monotonically increasing trend. This indicated that higher values in these feature topics positively influence satisfaction, aligning with findings from other studies. For CSS, SHAP values remained stable between -1 and 1 when sentiment values ranged from 0 to 0.4, reflecting relatively strong negative sentiment. As sentiment values increased from 0.4 to 0.8, SHAP values exhibited a monotonic rise, stabilizing beyond 0.8. For PE, SHAP values decreased when sentiment values were below 0.4, increased between 0.4 and 0.7, and stabilized above 0.7. For IME, SHAP values increased monotonically when sentiment values were below 0.6 and stabilized above 0.6.

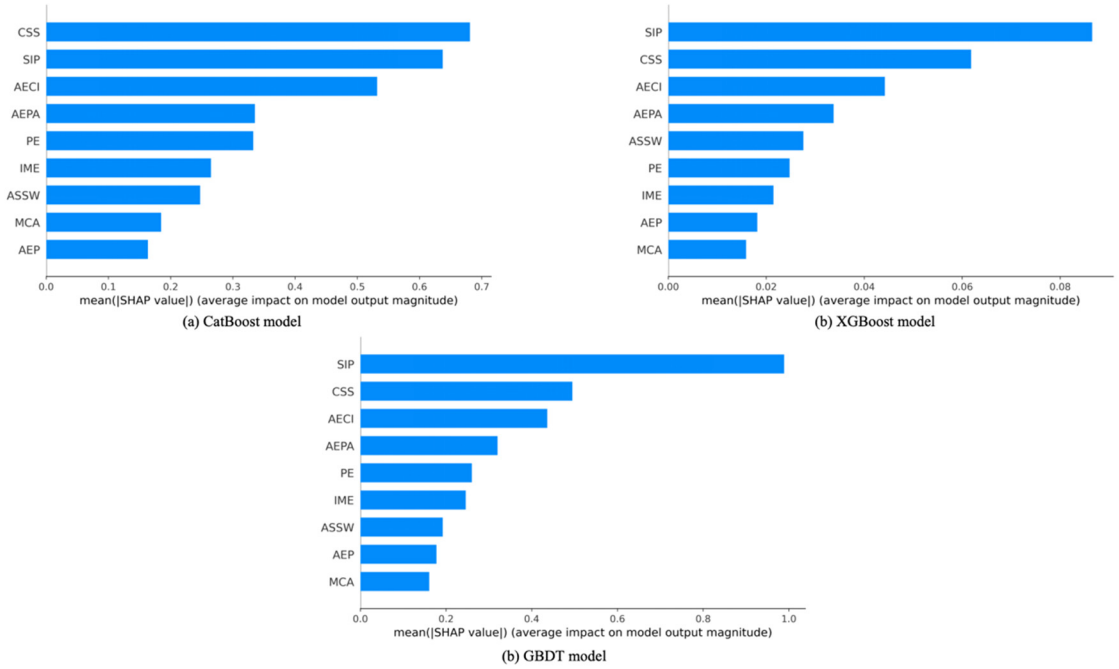

Figure 4. Average impact on model output magnitude.

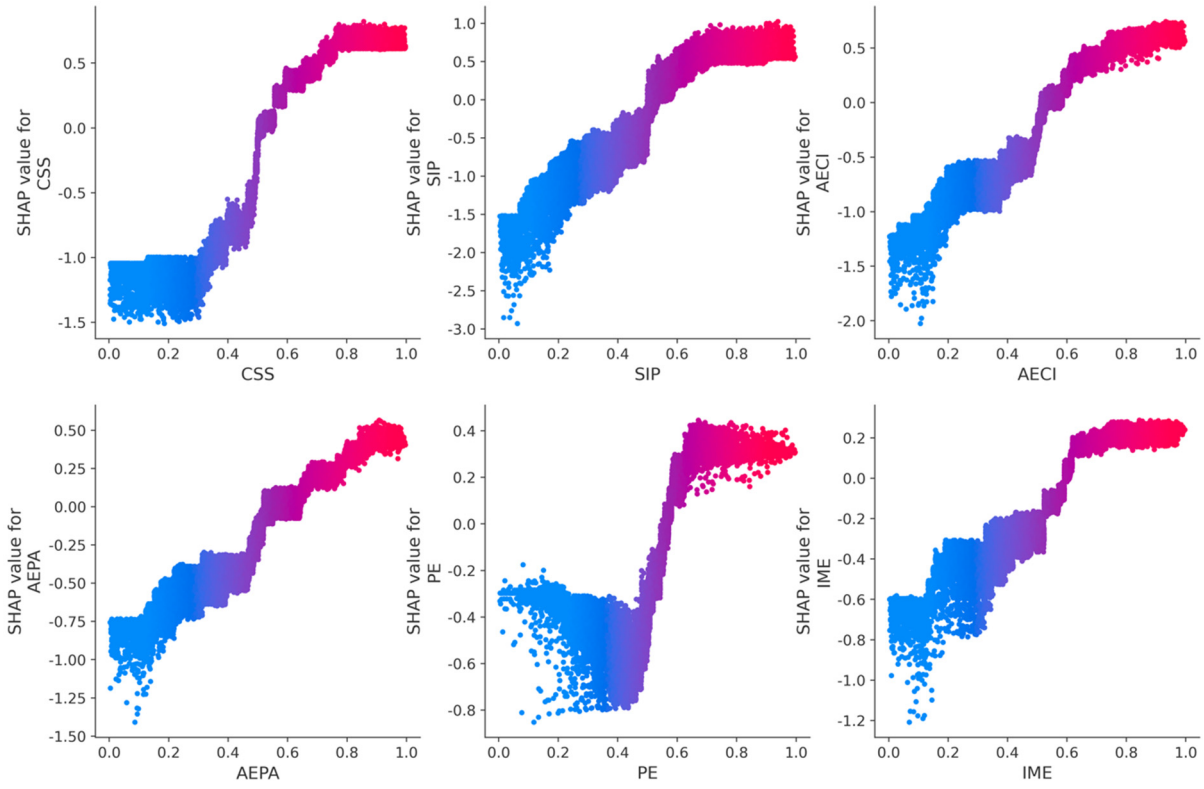

Figure 5. Influence of SHAP value on each feature.

4.5. Customer segmentation

Based on the SHAP method, the seven key determinants, CSS, SIP, AECI, AEPA, PE, IME, and ASSW, were identified based on feature importance. These determinants served as the basis for subsequent customer segmentation. The DEC algorithm is applied to segment users based on their fine-grained sentiment scores for these seven determinants. The quality of the clustering results is then evaluated using the SC (silhouette coefficient), a widely adopted metric that jointly considers intra-cluster cohesion and inter-cluster separation to quantify the appropriateness of cluster assignments. Conceptually, the SC captures the balance between within-cluster and between-cluster distances, where higher values indicate a more distinct and meaningful clustering structure. Following this principle, SC values are calculated for different numbers of clusters, as shown in Figure 6. Considering the practical requirements of e-commerce customer segmentation and the SC evaluation results, the optimal number of clusters was eight.

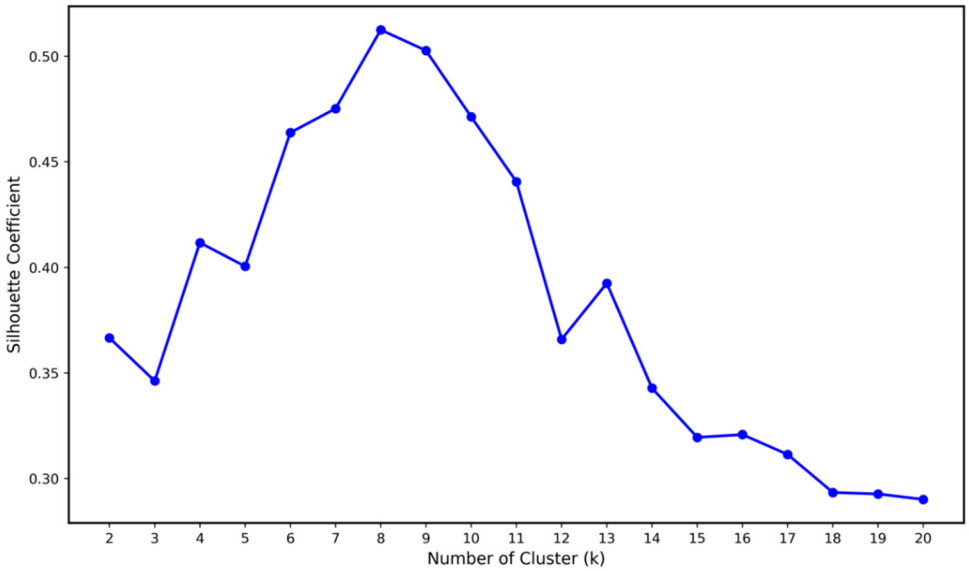

Figure 6. Silhouette coefficient line chart based on deep embedding clustering.

Table 5. Detailed results of the clusters from the DEC algorithm.

Cluster lable	Size	Population proportion(%)	Average rating
Customer segmentation 0	1517	6.80	3.82
Customer segmentation 1	3973	17.80	4.78
Customer segmentation 2	2833	12.69	2.51
Customer segmentation 3	3736	16.74	3.91
Customer segmentation 4	1633	7.32	4.64
Customer segmentation 5	2233	10.00	4.73
Customer segmentation 6	2209	9.90	1.62
Customer segmentation 7	4186	18.75	4.94

The clustering results are summarized in Table 5. Among them, Customer segmentation 7 and 1 contained the largest proportions of users, each exceeding 15%, and showed the highest average star

ratings, reflecting high satisfaction with products and services. Thus, these groups can be considered high-satisfaction clusters. In contrast, other customer segments, such as Customer segmentation 6, which represents the most distinctive user group, had the lowest average star ratings and were likely to belong to a group with pronounced emotional characteristics. This group represents a key target for platform merchants aiming to implement targeted marketing strategies and ensure service guarantees.

4.6. Customer profiling

Next, the clustering results are interpreted by analyzing the fine-grained sentiment tendencies of user reviews associated with each topic. Visual analyses of the fine-grained sentiment tendencies in each user cluster facilitate the development of targeted marketing strategies. As shown in Figure 7, the bar chart illustrates the distribution of sentiment tendencies across user clusters for each topic document. The bar length represents the average sentiment score for each feature topic, while color indicates sentiment polarity, with the saturation reflecting the intensity of sentiment.

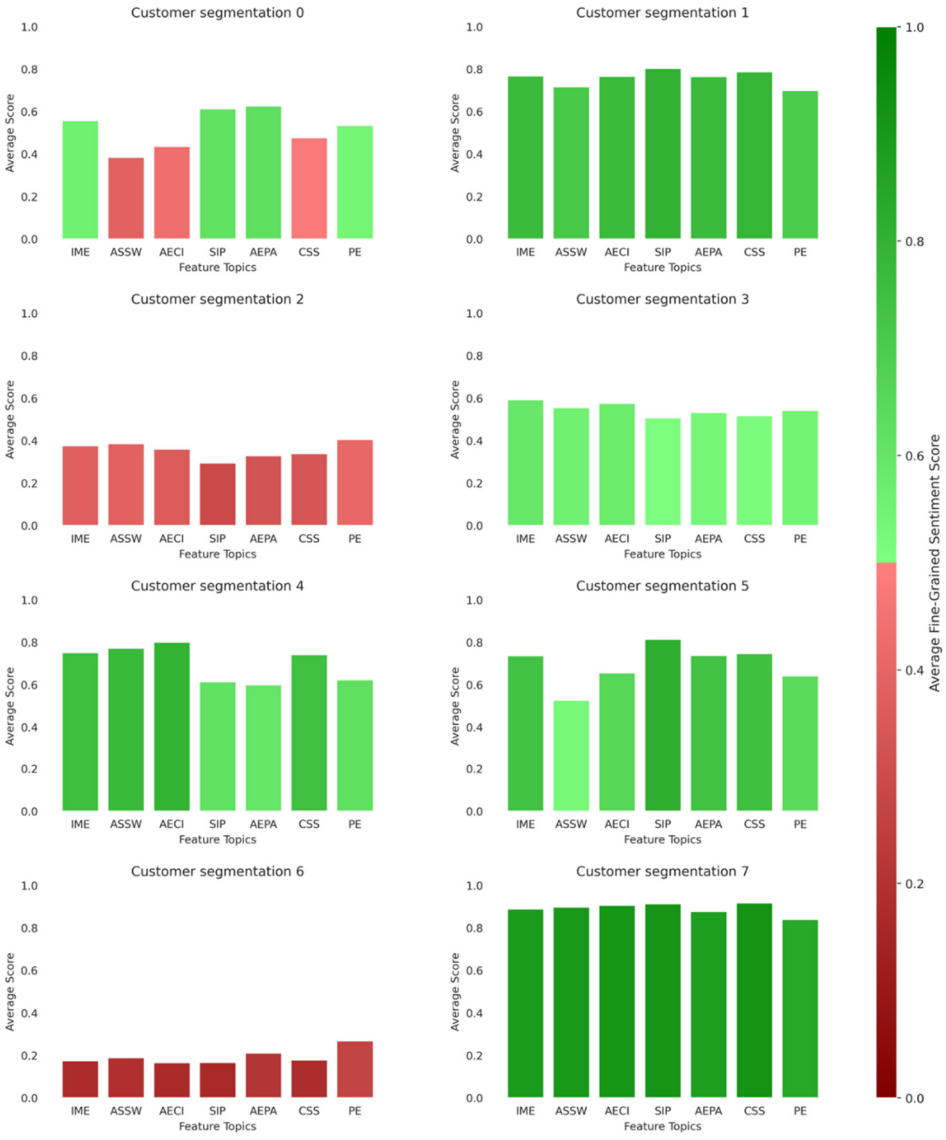

Figure 7. Fine-grained sentiment tendencies under each feature topic.

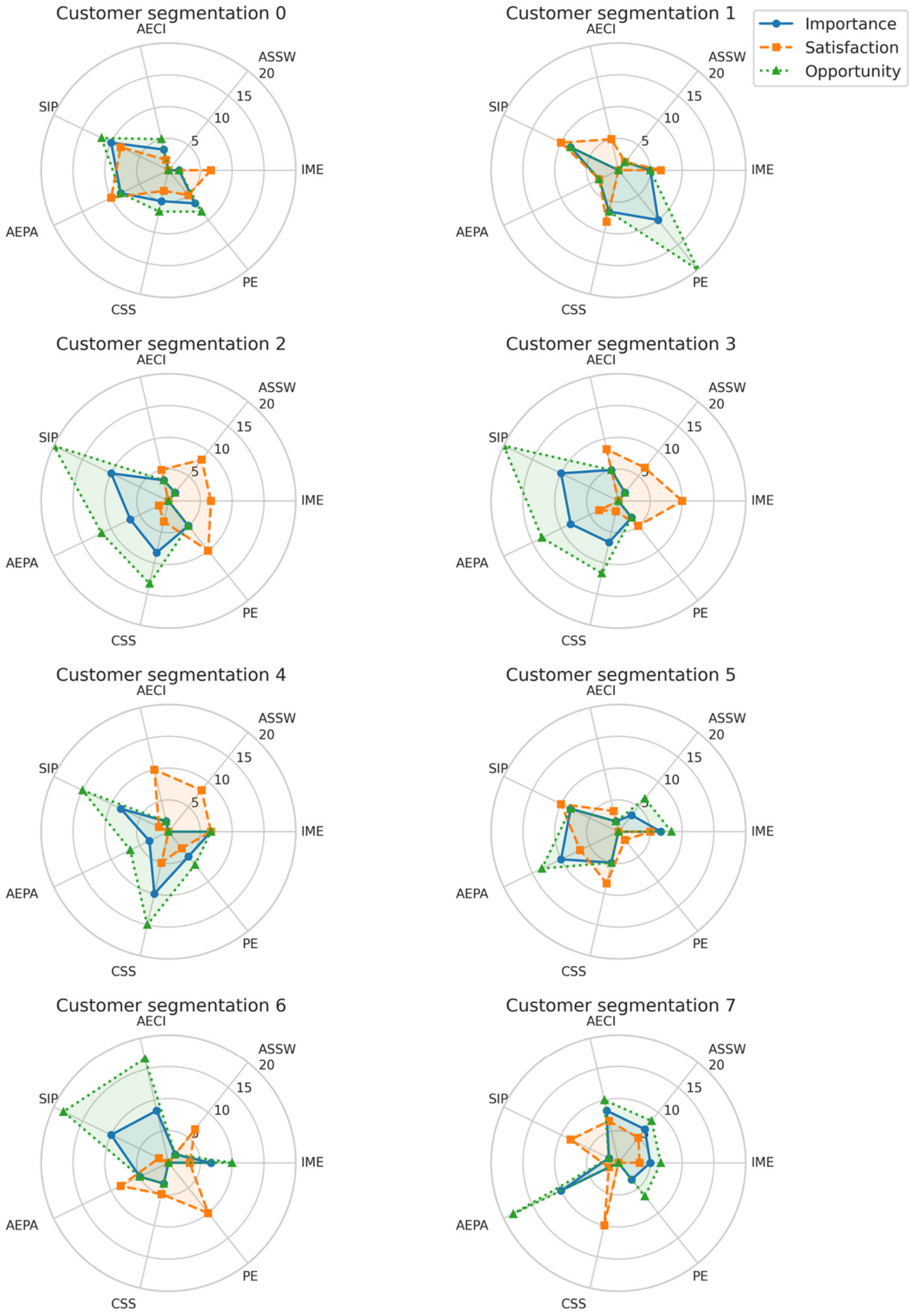

Figure 8. Opportunity scores of each customer segmentation.

For instance, in Customer segmentation 0, Figure 7 illustrates the sentiment tendencies across nine feature topics. Positive sentiment predominated in the topics of IME, SIP, AEPA, and PE. In contrast, negative sentiment was more prevalent in ASSW, AECI, and CSS. Overall, for users in Customer segmentation 0, it would be advantageous to continue promoting products that meet their expectations for quality and functionality. The negative feedback on after-sales service, audio equipment connections, and customer service may stem from unmet expectations regarding service, support, or poor connectivity with audio equipment. Therefore, enhancing the quality of audio equipment connections and interfaces, as well as improving customer and after-sales service, is recommended for these users.

In Customer segmentation 6, sentiment tendencies across all nine topic attributes were predominantly negative, making this group a key target for investigating product and service deficiencies, as well as monitoring after-sales service. This cluster exhibited the most frequent and intense negative sentiment in AECI, SIP, and CSS. Therefore, platforms and merchants should prioritize enhancing the quality of audio equipment connections, string instrument parts, and customer service. Furthermore, this group of e-commerce users should be closely monitored with respect to after-sales support for these product categories. Offering appropriate explanations and compensation could help mitigate dissatisfaction, reduce customer attrition, and maintain positive customer relations.

Opportunity scores are calculated based on the importance and satisfaction levels of product features for each customer segment (Figure 8). For Customer segmentation 6, SIP and AECI are product features that significantly impact overall customer satisfaction but exhibited lower satisfaction levels, indicating potential market opportunities. Thus, musical instrument companies could adopt a divide-and-conquer strategy to plan the development of next-generation products tailored to each customer segment. For instance, developing product versions with enhanced string instrument components and optimizing audio equipment connectivity and interface design to better meet user requirements could be effective strategies.

5. Discussions

To address the issue of information overload in e-commerce reviews, we develop a novel comprehensive method for customer segmentation based on identifying topics and sentiments from unstructured online product reviews. This enables the identification of potential user groups and the creation of decision profiles to support platform vendors in targeted marketing. In this section, we discuss the theoretical and practical implications of the proposed method, outline its limitations, and propose avenues for future research.

5.1. Theoretical implications

The demand-based customer segmentation discussed here is closely related to the development of new product concepts. In these studies, constructing the ML model with high accuracy and interpretability is crucial for explaining social phenomena [53]. However, in demand-based customer segmentation, there is no consensus on the importance and sentiment scores of product features used to characterize customers, as customers express their satisfaction or dissatisfaction with these features in diverse ways. In this study, customer preferences are estimated based on the fine-grained sentiment and importance values of product features derived from online reviews. To analyze the fine-grained

sentiment of features, a pre-trained BERT-base-uncased-sentiment model is used at the clause level, with TF-IDF word weights innovatively introduced to identify the sentiment scores of product features, ensuring that the results reflect the true sentiments expressed in the review texts. For importance estimation, the SHAP method is used, an IML technique that identifies and explains the nonlinear relationships between product feature satisfaction and overall customer satisfaction. This provides interpretive value for understanding how satisfaction with each product feature influences overall customer satisfaction.

In summary, this study contributes to new product development by providing a method for customer segmentation based on online product reviews. This research is distinguished from others by its focus on integrating feature topic extraction with identifying the fine-grained sentiment of the extracted feature topics within online product reviews, as well as using deep embedding clustering techniques to characterize each customer. Based on the sentiment and importance of these product features, a new method for customer segmentation has been developed.

5.2. Practical implications

The proposed method offers product development and design managers valuable insights into potential markets and new product concepts for customer-oriented development. By applying the proposed approach, product development and design managers can explore potential markets by identifying customer segments with high importance for specific product features but low satisfaction. They can also identify new product concepts for potential customers by enhancing the value of product features for each customer segment. Furthermore, the method includes identifying product features from online product reviews and estimating the sentiment and importance values of these features for customer segmentation. Each process provides customer demand information to product development and design managers via online reviews. In identifying product features, product development and design managers can determine which features customers are most interested in. When assessing the importance of product features, managers can identify those that are relatively important across all customers. Finally, the proposed method enables customer segmentation using online reviews, enabling product development and design managers to quickly collect a large volume of reviews at minimal cost.

Managers can leverage these factors to tailor and improve specific features, ensuring alignment with user preferences. Proactive service recovery measures can also be implemented to promptly address issues related to key factors, minimize dissatisfaction, and prevent negative reviews. Due to limited budgets and resources, managers may not address all USAT factors simultaneously. Therefore, prioritizing these factors to optimize applications and enhance user experience is crucial. Strategies to improve USAT should prioritize these factors to ensure they are robust, user-friendly, and aligned with user expectations. Investing resources and effort into enhancing these aspects may increase USAT and promote continued usage. This approach aids in competitive analysis, offering businesses a comprehensive understanding of their market's strengths, weaknesses, opportunities, and threats. Companies can use this information to make informed decisions on marketing strategies. Given the dynamic market, regular analyses are recommended to stay updated on changes.

5.3. Limitations and future research

This study has several limitations and directions for future improvement, which provide avenues for further research. First, the proposed method characterizes customers based on the importance values and sentiment scores of product features. In future research, researchers could incorporate additional demographic, psychological, and purchasing behavior data to identify customer segments with unmet needs [54]. Additionally, the temporal focus of this research may not fully capture the evolving nature of user sentiments and preferences, limiting the generalizability of the findings to other time periods. Longitudinal studies monitoring changes in USAT over time would provide valuable insights into the dynamic nature of user preferences and satisfaction determinants. In the future, researchers could explore combining NLP-based text mining with interviews from successful customers to examine the usefulness of reviews and their predictive factors. Finally, applying these techniques in multilingual environments remains an open question, warranting further exploration.

6. Conclusions

Based on unstructured online reviews, the innovative five-stage ensemble approach is proposed in this study, which integrates data collection and preprocessing, topic extraction modeling, topic sentiment vector modeling, determination of product features on star rating modeling, and customer segmentation based on topic sentiment score of product features to uncover and quantify users' fine-grained sentiments toward each product feature. To ensure a high degree of automation, the hyperparameters and thresholds are systematically specified for each stage of the proposed method. This approach retains fine-grained sentiment information about product features and addresses the lack of detailed sentiment data in many existing e-commerce clustering methods. Using the Word2Vec-based topic identification model, nine detailed determinants are identified as influential factors of USAT. Additionally, Clause-level sentiment analysis is conducted using a fine-tuned BERT model, in which sentiment scores are adjusted through TF-IDF-based weighting to enhance granularity. Thus, in this study, we conduct an extensive comparison of ten machine learning algorithms to identify the determinants associated with USAT. Among these, CatBoost outperforms all other evaluated models in accuracy and F1-score. By leveraging the IML method SHAP, we uncover the relative importance of determinants affecting USAT and highlight seven key factors. Additionally, we present an innovative application by integrating DEC with a topic-sentiment model and employing sentiment tendency visualization to construct e-commerce user decision profiles. Furthermore, opportunity scores are generated to assist managers in designing customized products by focusing on features that are highly important but exhibit low satisfaction, thereby facilitating personalized marketing across diverse customer segments.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Wang Y, Lu X, Tan Y, (2018) Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electron Commer Res Appl* 29: 1–11. https://doi.org/10.1016/j.elerap.2018.03.003
2. Park S, Kim H, (2024) Extracting product design guidance from online reviews: An explainable neural network-based approach. *Expert Syst Appl* 236: 121357. https://doi.org/10.1016/j.eswa.2023.121357
3. Kumar A, Chakraborty S, Bala PK, (2023) Text mining approach to explore determinants of grocery mobile app satisfaction using online customer reviews. *J Retail Consum Serv* 73: 103363. https://doi.org/10.1016/j.jretconser.2023.103363
4. Xuan Y, Zhang L, Bao H, Hu J, (2024) How to obtain product green design requirements based on sentiment analysis and topic analysis: Using washing machine online reviews as an example. *J Environ Manage* 365: 121454. https://doi.org/10.1016/j.jenvman.2024.121454
5. Alsayat A, (2023) Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Comput Appl* 35: 4701–4722. https://doi.org/10.1007/s00521-022-07992-x
6. Pal S, Biswas B, Gupta R, Kumar A, Gupta S, (2023) Exploring the factors that affect user experience in mobile-health applications: A text-mining and machine-learning approach. *J Bus Res* 156: 113484. https://doi.org/10.1016/j.jbusres.2022.113484
7. Joung J, Kim HM, (2021) Explainable neural network-based approach to Kano categorisation of product features from online reviews. *Int J Prod Res* 60: 7053–7073. https://doi.org/10.1080/00207543.2021.2000656
8. Bi JW, Liu Y, Fan ZP, Cambria E, (2019) Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *Int J Prod Res* 57: 7068–7088. https://doi.org/10.1080/00207543.2019.1574989
9. Zhang C, Zhang H, Wang J, (2018) Personalized restaurant recommendation method combining group correlations and customer preferences. *Inf Sci* 454: 128–143. https://doi.org/10.1016/j.ins.2018.04.061
10. Suryadi D, Kim HM, (2019) A data-driven methodology to construct customer choice sets using online data and customer reviews. *J Mech Des* 141: 111103. https://doi.org/10.1115/1.4044198
11. Rungruang C, Riyapan P, Intarasit A, Chuarkham K, Muangprathub J, (2024) RFM model customer segmentation based on hierarchical approach using FCA. *Expert Syst Appl* 237: 121449. https://doi.org/10.1016/j.eswa.2023.121449
12. Wang B, Miao Y, Zhao H, Jin J, Chen Y, (2016) A biclustering-based method for market segmentation using customer pain points. *Eng Appl Artif Intell* 47: 101–109. https://doi.org/10.1016/j.engappai.2015.06.005
13. Li Y, Meng C, Tian J, Fang Z, Cao H, (2024) Data-driven customer online shopping behavior analysis and personalized marketing strategy. *J Organ End User Comput* 36: 1–22. https://doi.org/10.4018/JOEUC.346230

14. Fang X, Zhou J, Pantelous AA, Lu W, (2024) A machine learning and clustering-based methodology for the identification of lead users and their needs from online communities. *Expert Syst Appl* 248: 123381. https://doi.org/10.1016/j.eswa.2024.123381
15. Guo L, Zhan J, Kou G, Martínez L, (2024) A sentiment analysis and dual trust relationship-based approach to large-scale group decision-making for online reviews: A case study of China Eastern Airlines. *Inf Sci* 667: 120515. https://doi.org/10.1016/j.ins.2024.120515
16. Harris ZS, (1954) Distributional structure. *WORD* 10: 146–162. https://doi.org/10.1080/00437956.1954.11659520
17. Blei DM, Ng AY, Jordan MI, (2003) Latent dirichlet allocation. *J Mach Learn Res* 3: 993–1022
18. Srivastava A, Sutton C, (2017) Autoencoding variational inference for topic models. Preprint, arXiv:1703.01488. https://doi.org/10.48550/arXiv.1703.01488
19. Ke J, Wang Y, Fan M, Chen X, Zhang W, Gou J, (2024) Discovering e-commerce user groups from online comments: An emotional correlation analysis-based clustering method. *Comput Electr Eng* 113: 109035. https://doi.org/10.1016/j.compeleceng.2023.109035
20. Yamarthi S, Chintala B, Rambabu R, Rao BY, Rao PV, Basha PH, (2025) Sentiment analysis framework for entropy-based product recommendation system. *Knowl Inf Syst* 67: 11611–11631. https://doi.org/10.1007/s10115-025-02570-8
21. Jeong B, Yoon J, Lee JM, (2019) Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *Int J Inf Manage* 48: 280–290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009
22. Xiao Y, Li C, Thürer M, Liu Y, Qu T, (2022) User preference mining based on fine-grained sentiment analysis. *J Retailing Consum Serv* 68: 103013. https://doi.org/10.1016/j.jretconser.2022.103013
23. Wang X, Goh DHL, (2020) Components of game experience: An automatic text analysis of online reviews. *Entertain Comput* 33: 100338. https://doi.org/10.1016/j.entcom.2019.100338
24. Liu Y, You TH, Zou J, Cao BB, (2024) Modelling customer requirement for mobile games based on online reviews using BW-CNN and S-Kano models. *Expert Syst Appl* 258: 125142. https://doi.org/10.1016/j.eswa.2024.125142
25. Park J, (2023) Combined text-mining/DEA method for measuring level of customer satisfaction from online reviews. *Expert Syst Appl* 232: 120767. https://doi.org/10.1016/j.eswa.2023.120767
26. Kumar A, Bala PK, Chakraborty S, Behera RK, (2024) Exploring antecedents impacting user satisfaction with voice assistant app: A text mining-based analysis on Alexa services. *J Retailing Consum Serv* 76: 103586. https://doi.org/10.1016/j.jretconser.2023.103586
27. Matzler K, Bailom F, Hinterhuber HH, Renzl B, Pichler J, (2004) The asymmetric relationship between attribute-level performance and overall customer satisfaction: A reconsideration of the importance-performance analysis. *Ind Mark Manage* 33: 271–277. https://doi.org/10.1016/S0019-8501(03)00055-5
28. Brin S, Page L, (2012) Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput Networks* 56: 3825–3833. https://doi.org/10.1016/j.comnet.2012.10.007
29. Boyd-Graber J, Mimno D, Newman D, (2014) Care and feeding of topic models: Problems, diagnostics, and improvements, In: *Handbook of Mixed Membership Models and Their Applications*, Chapman and Hall/CRC, 30.
30. Joung J, Kim HM, (2021) Approach for importance–performance analysis of product attributes from online reviews. *J Mech Des* 143: 081705. https://doi.org/10.1115/1.4049865

31. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, (2013) Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 3111–3119.
32. Xu X, Li Y, (2016) The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *Int J Hospitality Manag*e 55: 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003
33. Salton G, Buckley C, (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 24: 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
34. Zaki Ahmed A, Rodríguez-Díaz M, (2020) Significant labels in sentiment analysis of online customer reviews of airlines. *Sustainability* 12: 8683. https://doi.org/10.3390/su12208683
35. Chen T, Guestrin C, (2016) XGBoost: A scalable tree boosting system, In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
36. Breiman L, (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324
37. Freund Y, Schapire RE, (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 55: 119–139. https://doi.org/10.1006/jcss.1997.1504
38. Friedman JH, (2001) Greedy function approximation: A gradient boosting machine. *Ann Stat* 29: 1189–1232. https://doi.org/10.1214/aos/1013203451
39. Breiman L, (1996) Bagging predictors. *Mach Learn* 24: 123–140. https://doi.org/10.1007/BF00058655
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12: 2825–2830.
41. Bishop CM, (2006) *Pattern Recognition and Machine Learning*, New York: Springer.
42. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30: 3146–3154.
43. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A, (2018) CatBoost: Unbiased boosting with categorical features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6639–6649.
44. Cortes C, Vapnik V, (1995) Support-vector networks. *Mach Learn* 20: 273–297. https://doi.org/10.1007/BF00994018
45. Molnar C, Casalicchio G, Bischl B, (2020) Interpretable machine learning—a brief history, state-of-the-art and challenges, In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417–431.
46. Lundberg SM, Lee SI, (2017) A unified approach to interpreting model predictions, In: *Advances in Neural Information Processing Systems*, 4765–4774.
47. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2: 56–67. https://doi.org/10.1038/s42256-019-0138-9
48. Liu C, Li Y, Fang M, Liu F, (2023) Using machine learning to explore the determinants of service satisfaction with online healthcare platforms during the COVID-19 pandemic. *Serv Bus* 17: 449–476. https://doi.org/10.1007/s11628-023-00535-x
49. Xie J, Girshick R, Farhadi A, (2016) Unsupervised deep embedding for clustering analysis. In: *International Conference on Machine Learning*, 478–487.

50. Tao WB, Qian YR, Zhang YY, Ma HZ, Leng HY, Ma MN, (2022) Survey of deep clustering algorithm based on autoencoder. *Comput Eng Appl* 58: 16–25. https://doi.org/10.3778/j.issn.1002-8331.2204-0049
51. Ulwick AW, (2002) Turn customer input into innovation. *Harv Bus Rev* 80: 91–97.
52. Davies DL, Bouldin DW, (1979) A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 2009: 224–227. https://doi.org/10.1109/TPAMI.1979.4766909
53. Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T et al. (2021) Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int J Inf Manage* 57: 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
54. Zuo M, Angelopoulos S, Liang Z, Ou CX, (2023) Blazing the trail: Considering browsing path dependence in online service response strategy. *Inf Syst Front* 25: 1605–1619. https://doi.org/10.1007/s10796-022-10311-3

AIMS Press