

A Framework for Customer Needs Analysis: Integrating Dynamic Sentiment Lexicon Construction with Kano Model

Qiong He

Beijing Information Science & Technology University

Yijia Li

Beijing Information Science & Technology University

Zhenwei Yang

Beijing Information Science & Technology University

Shuai Wang

Beijing Information Science & Technology University

Jian Zhang

zhangjian@bistu.edu.cn

Beijing Information Science & Technology University

Article

Keywords: Customer needs identification, Dynamic sentiment lexicon, Kano model, Product improvement

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9432859/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Framework for Customer Needs Analysis: Integrating Dynamic Sentiment Lexicon Construction with Kano Model

Qiong He ¹, Yijia Li ², Zhenwei Yang ³, Shuai Wang ⁴, Jian Zhang^{5,*}

¹College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, 102206, China; e-mail: joanrun@126.com

²College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, 102206, China; e-mail: 17302261314@163.com

³College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, 102206, China; e-mail: 13618605664@163.com

⁴College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, 102206, China; e-mail: wangshuai202407@163.com

⁵College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, 102206, China; e-mail: zhangjian@bistu.edu.cn

Corresponding Author: Jian Zhang; e-mail: zhangjian@bistu.edu.cn; Telephone: +86- 13910320041

Abstract

Accurately analyzing customer needs is crucial for effective product design and improvement. Online reviews are a rich source of user feedback, but traditional static sentiment lexicons struggle to capture context-dependent emotional expressions in user-generated content (UGC). The traditional Kano model also relies heavily on subjective questionnaires, limiting its objectivity and scalability. To address these limitations, we propose an integrated customer needs analysis framework that combines dynamic sentiment lexicon construction with the Kano model. First, we use guided BERTopic to extract key product attributes from online reviews. Then, we build a dynamic domain-adapted sentiment lexicon using the TF-IDF-PageRank algorithm and BERT contextual embeddings. The Kano model classifies customer needs, and an opportunity algorithm ranks improvement priorities. This framework was evaluated based on 8,970 new energy vehicle reviews, identifying nine key attributes and quantifying actionable improvement priorities. This data-driven approach enhances the objectivity of needs analysis and provides a repeatable and transferable solution for

intelligent product optimization.

Keywords: Customer needs identification; Dynamic sentiment lexicon; Kano model; Product improvement

1. Introduction

Effective product design and iterative optimization rely on precise, data-driven identification of customer needs. With the popularity of e-commerce platforms and user-generated content, online reviews have become a low-cost and abundant source for obtaining authentic user preferences and emotional feedback. However, traditional customer needs analysis faces two key limitations: static general sentiment lexicons cannot adapt to domain-specific and context-dependent emotional expressions in UGC; and the conventional Kano model heavily depends on manual surveys, resulting in high subjectivity, small sample bias, and low efficiency. To overcome these challenges, this study develops an integrated customer needs analysis framework that combines dynamic sentiment lexicon construction with the Kano model. This framework automates attribute extraction, sentiment quantification, needs classification, and improvement priority ranking, providing an objective and scalable solution for data-driven product enhancement.

Product design and development constitute the core link in a market demand-driven process. Designers and developers leverage market demand insights to create products that align with consumer expectations. In order to gain advantages in increasingly competitive markets, manufacturers must constantly innovate, develop, and improve products to enhance market share [1].

In the product development process, a comprehensive understanding of customer needs is paramount to ensuring that products align with user expectations. However, traditional methods for collecting customer demands—such as questionnaires, telephone interviews, and brainstorming—often encounter significant limitations. These include restricted sample sizes, response lag, and subjective data interpretation, which can compromise demand capture accuracy and lead to biased decision-making. With the proliferation of e-commerce platforms, community websites, and third-party

review platforms [2,3], online reviews have emerged as a valuable data source for discerning customer needs across various sectors, including hotels [4], airlines [5], and restaurants [6]. These reviews contain rich information pertaining to consumer preferences, requirements, and sentiments towards products. Furthermore, researchers have leveraged traditional quality tools—such as Quality Function Deployment (QFD), Failure Mode and Effects Analysis (FMEA), and the Kano model—to enhance product quality. By systematically identifying user needs, preempting potential failures, and optimizing design priorities, these structured approaches support product improvement efforts [7–9].

By adopting a data-driven approach, enterprises can leverage the rich user-generated content to conduct in-depth exploration and sophisticated analysis of market demands, thereby enabling more precise product design and enhancement. This methodology not only elevates the efficiency and accuracy of product development but also empowers businesses to adapt more swiftly to market dynamics, significantly bolstering product competitiveness.

2. Literature Review

2.1 Customer needs analysis

Kano model is a tool to explore the relationship between customer needs and satisfaction. According to it, needs are divided into different types, such as charm, expectation, necessity, reverse and indifference [10]. Traditionally, the Kano model serves as a qualitative analysis tool, integrating the Kano questionnaire, Kano evaluation table and Kano survey result table, constitute a set of systematic methods [11]. Through structured user interviews and surveys, this method effectively captures and analyzes the deep and implicit cognition, expectation and satisfaction evaluation of customers for specific products or services [12]. Baier et al. (2020) conducted a survey on existing customers and potential customers of a large sporting goods retailer and systematically evaluated 37 service improvement schemes, which provided an empirical basis for multi-channel retailers to improve customer shopping satisfaction [13]. Chen et al. (2021) systematically evaluated passengers' satisfaction with transportation service

elements using the Kano model, and conducted differentiated analysis based on passenger types and transportation methods to provide priority strategies for improving core services [14]. Sinemus et al., (2025) employed the Kano model to conduct an in-depth investigation into the impact of 25 shopping application features on consumer satisfaction, while identifying the demand variations among distinct user segmentation groups [15]. However, although the Kano model shows strong qualitative analysis ability in its practical application, it is also faced with the problems of strong subjectivity of classification methods and insufficient response to customers' complex psychology [16]. Moreover, the evaluation results of this model are highly dependent on the selected survey samples, which may be affected by sample selection bias.

In order to overcome these limitations, some scholars try to use online reviews as the data source of Kano model, and enhance the quantitative analysis and objectivity of Kano model through data-driven methods. Zhang and Luo (2022) takes online review data as material, conducts sentiment analysis on review content to measure the satisfaction degree of product attribute demand, and uses regression model to divide consumers' demand for product attributes into five types in Kano model [10]. Bi et.al (2019) extracted key product attributes from online reviews, using SVM model quantitative consumer emotional tendency of product attributes, construction of neural network model to analyze the impact on consumer satisfaction product attributes, on this basis, based on the Kano model will be divided into different consumer demand for product attribute category [17]. Zhao et.al (2020) waiting in the heart of the research on fuzzy theory combined with the Kano model, classifying extracted online reviews of the user requirements and the emotion analysis, finally constructs the demand - satisfaction quantification model, implements the quantitative evaluation of users' satisfaction [18]. By deeply mining the demand information implied in the review data, enterprises can not only obtain the opportunity to understand the feedback of customers, but also master the relevant intelligence of competitors, so as to optimize the attributes of their products or services and create commercial value.

Kano model is a tool to effectively explore the relationship between customer needs and satisfaction, and helps enterprises accurately identify and optimize product

attributes by dividing needs into different types. Combined with online review and other data sources, this model can overcome the limitations of traditional methods and achieve more objective and quantitative demand analysis, thus improving customer satisfaction and competitiveness, and creating business value for enterprises.

2.2 Sentiment Orientation Analysis

In the field of sentiment analysis, the present approach can be mainly divided into two classes: based on machine learning methods and emotional dictionary method.

Machine learning-based methods rely on extracting effective sentiment features from manually annotated datasets concentrated extract effective features of emotional tendency to train classification model based on these characteristics, in order to realize the text emotional polarity judgment. This process involves the complex characteristics of the engineering, such as vocabulary, syntax and semantic characteristics of multi-dimensional extract, as well as selecting or designing appropriate classifiers such as SVM [19], Naive Bayes [20], LSTM [21], deep learning models such as CNN [22], RNN [23] for model training and sentiment classification. There are also a series of improved models, such as IOVO-SVM [4], N-gram-NB [24], Bi-LSTM [25], TextCNN [26]. Although this method is able to deal with complex language phenomena and emotional expressions, its performance is highly dependent on the quality of labeled data and the rationality of feature selection.

Method based on emotional dictionary use pre-built emotional dictionary, by matching the words in the text, and the emotional words in the dictionary, thus computed text emotional polarity and emotional intensity. Emotional dictionary construction is the key of this method, it can be arranged through artificial, automatically extract, or a combination of both. Sentiment lexicon can be divided into two categories: domain sentiment lexicon and general sentiment lexicon. At present, the widely used general sentiment dictionaries include Taiwan University Sentiment Dictionary NTUSD [27], Dalian University of Technology Sentiment Vocabulary Ontology Library [28], HowNet dictionary [29], and Buzzword dictionary [30]. The basic sentiment lexicon contains a large number of known sentiment words. However,

it is far from enough to only rely on the basic sentiment lexicon to identify the sentiment words in a sentence.

In the era of big data, the generation of new sentiment words is particularly rapid. More and more Internet users generate documents with certain emotions, and a large number of new sentiment words have been introduced to accurately express specific emotional states. New emotional words are either new words or words with completely different meanings, which are used to express a certain emotion in a specific field. Therefore, we need a new method to capture new emotional words. Based on the review corpus, Turney (2002) uses the difference between the point mutual information of a candidate opinion word and positive and negative seed words to measure the sentiment orientation of the word. The principle is that words that appear in similar patterns in the text have high semantic and sentiment similarity [31]. Wang et.al (2012) proposed a word polarity calculation method combining HowNet and PMI [32]. This method uses HowNet to expand synonyms, which effectively reduces the problem caused by the low frequency of sentiment words in the corpus. Experimental results show that the average performance of the proposed method is improved by 5%. The sentiment seeds are extracted from the user review data set, and the general sentiment words are obtained through the graph propagation algorithm. The syntactic and sequential patterns are automatically detected by using the general sentiment words and the target words in the domain-specific corpus. Yang et.al (2019) first extracts candidate sentiment words from the reviews, and then selects the top k positive and negative words from the reviews as the seed word list. The WordNet1 model is used to calculate the similarity and add synonyms to expand the seed words to form the final sentiment lexicon. The candidate words are matched with the words in the sentiment dictionary to extract the final opinion words [33].

In the domain of sentiment lexicon construction, despite the substantial advancements achieved by traditional statistical methods and semantic representation strategies, two critical bottlenecks remain. First, static word vectors (e.g., Word2Vec, GloVe) struggle to capture context-sensitive sentiment semantic shifts. Second, existing pattern-matching approaches exhibit limited generalization capabilities for emerging

sentiment expressions on the Web and are ill-suited to adapt to the rapid evolution of unstructured sentiment expressions. To address these limitations, researchers have recently begun exploring a novel paradigm that integrates deep pre-trained language models with statistical learning. By employing a domain-adaptive seed word selection algorithm in conjunction with BERT's (Bidirectional Encoder Representations from Transformers) dynamic contextual representation dictionary construction method, this approach achieves an end-to-end automated process. It effectively resolves the dual contradiction between the cost of manual annotation and domain transferability, offering a new technical pathway for real-time dictionary updates in UGC scenarios.

3. Research Methods

Extracting product attributes and identifying customer needs from online review data is accomplished through the following four-stage procedure: (1) text collection and preprocessing; (2) customer demand identification; (3) sentiment analysis of product attributes based on a sentiment lexicon; (4) customer demand classification and product improvement suggestions. Figure 1 depicts the framework of the proposed method.

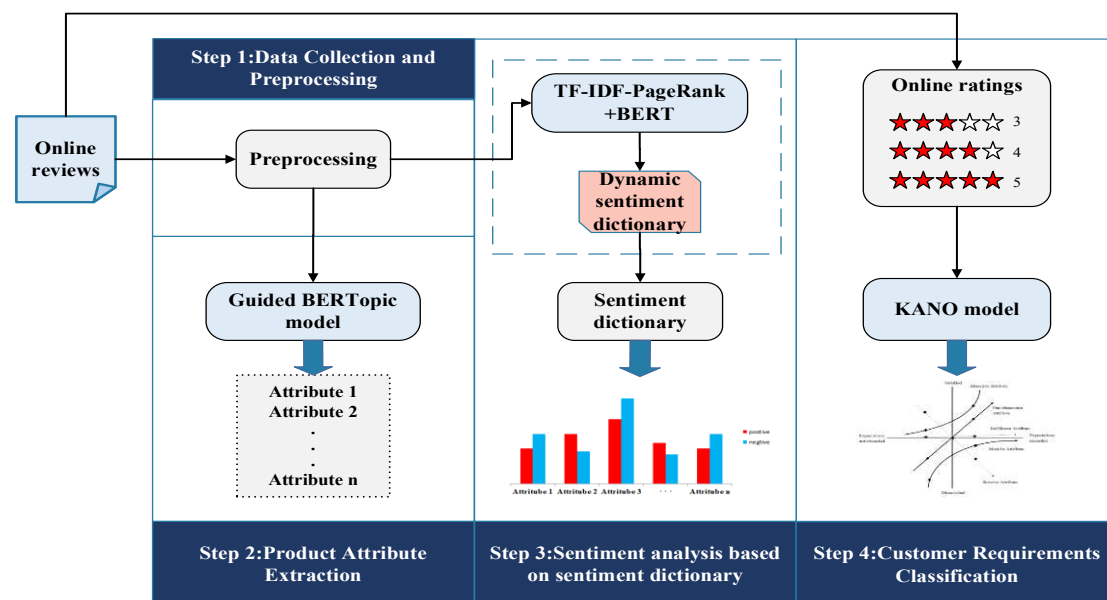


Figure 1 The framework of the proposed method

3.1 Framework Overview

First, product customer reviews are systematically collected from e-commerce

platforms. Second, in accordance with the principles of Natural Language Processing (NLP), the review texts undergo preprocessing. Specific operations encompass crawling rating scores and comment texts from online reviews, filtering out all non-Chinese characters and punctuation marks, followed by removing stop words and utilizing the widely adopted Chinese word segmentation toolkit Jieba for segmenting words and tagging parts of speech. Consequently, noise and low-information-content words are eliminated from the review texts, while valuable information-bearing words are retained.

3.2 Dynamic Sentiment Lexicon Construction

Guided BERTopic is a topic modeling approach that integrates domain-specific knowledge with data-driven methodologies. The essence of Guided BERTopic lies in guiding the topic generation process through predefined seed words. The purpose of incorporating the seed word guidance mechanism is to ensure that BERTopic focuses on topics relevant to this study, thereby avoiding the generation of irrelevant terms. When utilizing traditional Latent Dirichlet Allocation (LDA) and BERT models for topic clustering, issues such as insufficiently coherent topics and an abundance of noise words often arise. Furthermore, the incorporation of seed word guidance enables the identification of topics of high public concern while also assisting in uncovering latent concerns embedded within customer feedback. Built upon the BERT architecture, the model achieves semantic representation learning and incorporates a five-stage modeling procedure.

Step 1: Document Embedding: We employ the paraphrase-multilingual-MiniLM-L12-v2 model to map Chinese text into a high-dimensional vector space, thereby generating the semantic embedding vector e_d of the document.

Step 2: Seed Word Guidance: We introduce the seed word set S , which is defined by domain experts. This set guides the direction of constrained topic generation and facilitate knowledge fusion through the two steps:

(1) Seed Word Embedding Computation: For each seed word $w \in S$, we extract its BERT embedding vector $\text{BERT}(w)$, and subsequently compute the centroid vector

of the set:

$$e_s = \frac{1}{|S|} \sum_{w \in S} \text{BERT}(w)$$

Let $|S|$ denote the number of seed words, and let the centroid vector e_s characterize the semantic center of the seed word set.

(2) Semantic similarity guidance: The cosine similarity between the document embedding e_d and the seed word centroid e_s is computed, and a similarity threshold is established to enhance the relevance of documents related to the seed word in subsequent clustering.

Step 3: Dimensionality Reduction: The UMAP (Uniform Manifold Approximation and Projection) algorithm is employed to perform nonlinear dimensionality reduction on high-dimensional embedding vectors. UMAP not only enhances computational efficiency but also preserves both the local and global structures of the data, thereby providing low-dimensional, separable features for subsequent clustering tasks.

Step 4: Clustering: Based on the reduced feature space, the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm is utilized to conduct document grouping. This algorithm identifies clusters of arbitrary shapes through density-based clustering and automatically isolates noise points, effectively accommodating the sparsity and heterogeneity inherent in text data.

Step 5: Topic Keyword Distribution Generation: The improved c-TF-IDF (Class-based Term Frequency-Inverse Document Frequency) algorithm is applied to generate topic keyword distributions, as described by the following formula:

$$\text{c-TF-IDF} = \frac{f_{w,c}}{\sum f_{w,c}} \times \log \left(1 + \frac{N}{\sum_{c'} I(f_{w,c'} > 0)} \right)$$

The term $\frac{f_{w,c}}{\sum f_{w,c}}$ represents the relative frequency of term w within the theme c , where N is the total number of subjects. The expression $\sum_c I(f_{w,c} > 0)$ indicates the number of themes in which the word w appears. The formula $\log \left(1 + \frac{N}{\sum_{c'} I(f_{w,c'} > 0)} \right)$ measures the degree of differentiation of word w across subjects. Seed word enhancement strategy: A weight factor $\alpha > 1$ is applied to seed words, and their IDF

values are adjusted as follows:

$$\text{IDF}_{\text{guided}}(w) = \alpha \cdot \log \left(1 + \frac{N}{\sum I(f_{w,c'} > 0)} \right)$$

By enhancing the discrimination weight of seed words within a topic, their role as topic identifiers is significantly reinforced. During the Guided BERTopic modeling process, the incorporation of seed words serves as a guiding mechanism for constructing the topic model. Seed words, which are pre-selected keywords related to the topic, can influence the generation process of the topic model, encouraging it to converge more effectively toward the predefined topic category. This not only improves the consistency of the topic but also enhances its interpretability.

3.3 Step Guided BERTopic for Customer Need Extraction

Existing studies indicate that the application of general sentiment lexicons in professional domains is associated with significant semantic deviations [28]. These deviations primarily stem from three factors: the uniqueness of domain-specific sentiment words, the dynamic evolution of online language, and context sensitivity. Particularly in User-Generated Content (UGC) scenarios, the interplay between informal linguistic features and emerging Web semantics challenges the static thesauri of traditional dictionaries, rendering them inadequate for dynamic semantic representation. While manually annotated domain dictionaries can enhance emotion recognition accuracy, their high annotation costs and limited scalability constrain practical applications. In light of this, this study proposes an innovative framework for automatic construction that integrates statistical methods with deep learning techniques. Specifically, the framework employs the TF-IDF weighting mechanism and PageRank algorithm to identify highly discriminative sentiment seed words from domain corpora. Subsequently, the BERT model is utilized for context-aware word vector representation to perform semantic space dimensionality reduction. Lastly, a dynamic expansion mechanism based on cosine similarity is developed to facilitate the domain-adaptive expansion of the seed thesaurus.

3.3.1 TF-IDF-PageRank sentiment seed word selection

In this paper, the e-commerce sentiment lexicon developed by Soochow University is integrated with the Hownet sentiment lexicon to construct a fundamental positive and negative sentiment lexicon. Given the consistency between Soochow University's e-commerce sentiment dictionary and the text corpus selected for this study, as well as the general applicability of the Hownet sentiment dictionary in the Chinese domain, these two resources can complement each other to build a comprehensive sentiment lexicon with high coverage [32]. Traditional seed word selection methods suffer from limitations in single-feature modeling. While TF-IDF effectively quantifies the local statistical significance of words, it overlooks semantic correlations between terms. Although PageRank captures global structural characteristics of semantic association networks, its uniform initialization strategy struggles to emphasize domain-specific key terms. To address these issues, this paper proposes a TF-IDF-PageRank heterogeneous feature fusion model, with the following core steps:

First, the TF-IDF value of term t in document d is calculated using the TF-IDF algorithm.

$$\text{TF-IDF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log\left(\frac{N}{n_t + 1}\right)$$

Where N represents the total number of documents, n_t denotes the number of documents containing term t , and $f_{t,d}$ indicates the frequency of term t in document d .

Subsequently, the word co-occurrence network $G = (V, E, W)$ is constructed, where the vertex set V corresponds to the candidate words, the weight w_{ij} of the edge set E is determined based on the window co-occurrence frequency, and the importance of nodes is assessed using the PageRank algorithm:

$$PR(w_i) = \frac{1-d}{|V|} + d \sum_{w_j \in B(w_i)} \frac{PR(w_j)}{L(w_j)}$$

Where $d=0.85$ is the damping factor, $B(w_i)$ represents the set of incoming nodes for w_i , and $L(w_j)$ denotes the total number of outgoing links from w_j .

Unlike the uniform initialization in the traditional PageRank algorithm, the TF-IDF value is incorporated as a prior weight. Terms with high TF-IDF scores typically

convey domain-specific semantic information, and assigning higher initial weights can direct the PageRank algorithm to prioritize key concepts within the domain. Additionally, TF-IDF-based initialization enhances the convergence rate of the PageRank iteration process.

Finally, the final term importance score is derived by calculating the linearly weighted score:

$$S(w_i) = \alpha \cdot \text{TF-IDF}(w_i) + (1 - \alpha) \cdot \text{PR}(w_i), \alpha \in [0,1]$$

Where $\alpha=0.6$ is the optimal hyperparameter determined by grid search (search range: 0.1–0.9, step=0.1) on 10% validation set.

3.3.2 Word Vector Representation Leveraging the BERT Model

BERT overcomes the limitations of semantic representation in traditional word vector models by leveraging deep bidirectional context encoding and a large-scale pre-training followed by fine-tuning paradigm. Unlike traditional word vector models such as Word2Vec and GloVe, which are limited to generating static global vectors for each word, BERT employs the Transformer architecture. Through its Self-Attention mechanism, it enables bidirectional context encoding, thereby generating dynamic word representations that better capture contextual nuances.

$$h_i^{BERT} = f_{Transformer}(w_i | \text{Context}_{left}, \text{Context}_{right})$$

This mechanism enables the vector representations of the same word in different contexts to be distinctly differentiated, thereby accurately capturing contextual semantics.

3.3.3 Dynamic Expansion of seed Words based on cosine Similarity

Cosine Similarity is employed to quantify the semantic association between vectors:

$$\text{Sim}(v_{candidate}, v_{seed}) = \frac{v_{candidate} \cdot v_{seed}}{\|v_{candidate}\| \cdot \|v_{seed}\|}$$

Let $v_{candidate}$ denote the candidate word vector and v_{seed} represent the seed word vector. The cosine similarity threshold for seed word dynamic expansion is set to

0.75, which is determined by domain adaptability tests.

3.4 Customer requirement identification

The Kano model is a two-dimensional theoretical framework introduced by Japanese quality management expert Professor Noriaki Kano and his colleagues in 1984. It aims to elucidate the relationship between user satisfaction with products or services and the extent to which these offerings fulfill user needs [34]. Through analyzing variations in user satisfaction, the model categorizes five types of user requirements: attractive, one-dimensional, essential, reverse, and indifferent. As illustrated in FIG. 2, these five requirement types can be visualized in a two-dimensional space, providing a more intuitive understanding of the interconnection between user requirements and satisfaction levels.

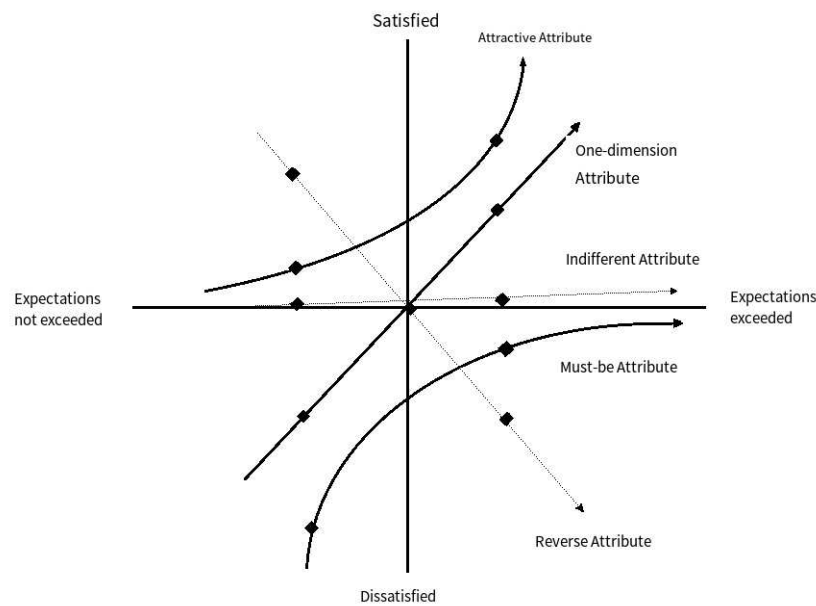


Figure 2 Kano model

(1) The attractive attribute pertains to unexpected demands that can surprise users. While its absence does not lead to dissatisfaction, its fulfillment significantly enhances user satisfaction.

(2) One-dimensional attributes refer to requirements that users typically expect and consider as a given. Their satisfaction exhibits a positive linear relationship with user satisfaction; as these needs are progressively met, user satisfaction increases accordingly, and conversely, decreases when unmet.

(3) Essential attributes represent basic needs deemed indispensable by users. Even when fully satisfied, they do not markedly improve user satisfaction.

(4) The reverse attribute denotes demands that users reject or resist. If such demands are not provided, users may remain satisfied or indifferent; however, their provision results in a negative linear decline in user satisfaction.

(5) The indifference attribute refers to demands with minimal impact on user satisfaction. Whether provided or not, these demands do not significantly alter user satisfaction levels. After analyzing customer sentiment for each product attribute and the associated review scores, the category of each product attribute can be determined based on the contribution of positive and negative sentiments.

Therefore, referring to the rules provided by Qi et al. (2016) [35], the formula of Kano model is as follows :

$$y = \alpha + \sum_{j=1}^n (\beta_j^{pos} X_j^{pos} + \beta_j^{neg} X_j^{neg}) \quad (3.1)$$

Where y represents customer satisfaction, and its value is the actual rating given by the customer for each review. X_j^{pos} and X_j^{neg} denote whether the customer has a positive or negative attitude toward product attribute j . When $X_j^{pos} = 1$, the customer has a positive attitude toward attribute j , and $X_j^{neg} = 1$, the customer has a negative attitude toward attribute j . β_j^{pos} represents the degree of customer's positive sentiment towards product attribute j , β_j^{neg} represents the degree of customer's negative sentiment towards product attribute j , and α is a constant term.

According to the mapping rules provided by Qi et al. (2016) [35], product attributes can be mapped to different categories.

The parameters can be estimated according to the formula. For each attribute, we have three observation points, which are $(-1, X_j^{neg})$, $(0,0)$ and $(1, X_j^{pos})$. As shown in Figure 4, through these three observation points, we can map attributes to different categories according to the Kano model. If the preference has been close to zero, $|\beta_j^{pos}| \leq \lambda$ and $|\beta_j^{neg}| \leq \lambda$, attributes can be mapped to different attributes, λ

is the threshold value must be set. Otherwise, if the three observations look like a straight line, $|\beta_j^{pos} + \beta_j^{neg}| \leq \delta * |\beta_j^{pos} - \beta_j^{neg}|$, when $\beta_j^{pos} + \beta_j^{neg} \geq 0$, the attribute can be mapped to a one-dimensional attribute, and when $\beta_j^{pos} + \beta_j^{neg} < 0$, the attribute can be mapped to a reverse attribute. Otherwise, when $\beta_j^{pos} + \beta_j^{neg} > 0$, the attribute can be mapped to the attractive attribute, and when $\beta_j^{pos} + \beta_j^{neg} < 0$, the attribute can be mapped to the necessary attribute.

Figure 3 shows the Kano model decision tree constructed according to the mapping rules.

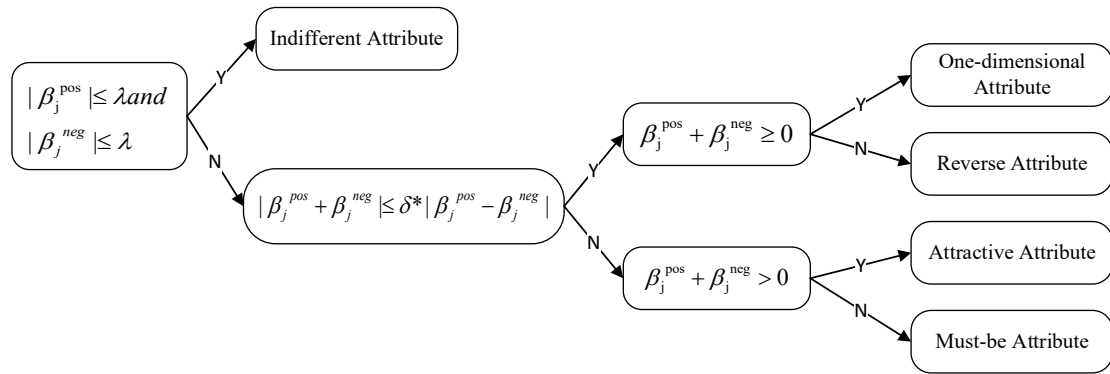


Figure 3 Kano model decision tree

Although the Kano model effectively categorizes each subject into distinct attribute categories, it exhibits limitations in prioritizing improvements. In this study, the opportunistic algorithm is integrated to extend the Kano model, thereby clarifying which unmet car attribute requirements should be addressed preferentially. The opportunistic algorithm calculates the chance score of each attribute based on two dimensions: importance and satisfaction. Importance reflects the user's focus on demand topic t , measured by the frequency of occurrence of the topic in the text. This frequency is calculated using the formula $CI_t = N_t/N$, where N_t represents the number of documents mentioning topic t within the review corpus, and N denotes the total number of documents. Satisfaction indicates the user's emotional inclination toward topic t and is derived by aggregating the emotional scores of related texts. The total emotional score is computed as $CS_t = \sum_{i=1}^t s_i^t (i = 1, 2, 3 \dots)$, where s_i^t represents the emotional score of the i th document belonging to topic t .

Consequently, the opportunistic algorithm is formulated as follows:

$$\begin{aligned} \text{Opportunity} &= \text{Importance} + \text{Max}(\text{Importance} - \text{Satisfaction}, 0) \\ &= \frac{CI_t - CI_{min}}{CI_{max} - CI_{min}} + \text{Max}\left(\frac{CI_t - CI_{min}}{CI_{max} - CI_{min}} - \frac{CS_t - CS_{min}}{CS_{max} - CS_{min}}, 0\right) \end{aligned}$$

$\frac{CI_t - CI_{min}}{CI_{max} - CI_{min}}$ and $\frac{CS_t - CS_{min}}{CS_{max} - CS_{min}}$ are used to keep the dimensions of importance and

satisfaction consistent.

4. Results

This section presents the objective results of customer needs analysis for new energy vehicles (NEVs) using the proposed framework, including data collection statistics, product attribute extraction outcomes, dynamic sentiment lexicon construction results, Kano model classification outputs, and product improvement priority rankings.

4.1 Data Collection

Dongchedi is a prominent automobile information service platform in China, dedicated to providing users with comprehensive automobile information, car purchase guidance, owner reviews, and video evaluations. The Dongchedi platform encompasses an extensive range of car brands and models, featuring a wealth of professional analyses and authentic user feedback. Consequently, we selected the Dongchedi website as the source for collecting our online review dataset. Given that new energy vehicles (NEVs) offer substantial advantages in terms of environmental protection, high energy efficiency, and government policy support, they have increasingly become a favored choice among consumers. Therefore, we chose NEVs as the focus of our research. We systematically gathered customer reviews from the five most popular NEV brands---Tesla, BYD, Chang'an, NIO, and Lixiang---that were posted on Dongchedi between February 2023 and December 2025. After data cleaning (removing duplicates, filtering out reviews with fewer than 10 Chinese characters, and excluding invalid entries such as default text or advertisements), 8,970 valid reviews were retained for analysis.

4.2 Product attribute extraction

A total of 30 preliminary topics were generated, and 9 core NEV product attributes were finally

obtained after topic merging and deduplication. The high-frequency keywords and corresponding attribute definitions of each core topic are listed in Table 1.

Table 1 Product attributes extracted from product reviews

Variable	Topic	High-frequency Keywords
X_1	Driving range	kilometers, battery, power, Driving range, sub-zero, wear and tear
X_2	Space	spacious, compact, trunk, legroom, large, second-row space
X_3	Appearance	model, body, design, fashionable, atmosphere, simplicity, face value
X_4	Intelligence	intelligent driving, assist system, navigation, wireless charging, automatic parking
X_5	Interior	good-looking, dashboard, odor, texture, material, display screen
X_6	Price	same price, cost-performance, discounts, maintenance costs, subsidies
X_7	Seats	heating, ventilation, noise, adjustment, leather
X_8	Chassis system	chassis, bearing, radar, tire pressure, noise
X_9	Driving experience	acceleration, pushback, powerful, responsive, Performance

4.3 Sentiment Analysis

For the construction of the domain sentiment lexicon, first, based on the foundational sentiment lexicon, the top 500 positive and negative seed sentiment words were selected using the TF-IDF-PageRank algorithm. Subsequently, the BERT model was employed to characterize the words in the corpus. A combination of cosine similarity and a dynamic weighting strategy was utilized to determine whether a candidate word belonged to the category of positive or negative sentiment words. If the candidate word was identified as a positive sentiment word, it was incorporated into the positive sentiment dictionary; if it was identified as a negative sentiment word, it was added to the negative sentiment dictionary. The final domain sentiment lexicon comprises 736 sentiment words, including 512 positive sentiment words and 224 negative sentiment words. Selected results from the sentiment lexicon are presented in Table 2:

Table 2 Sample sentiment lexicon words

Emotional orientation	Examples
Positive	Tenacity, sense of movement, novelty, roundness, fashion, refreshing, strong, sharp corners, steady, atmosphere, small, excellent performance, unique, simple, three-dimensional, reasonable layout, avant-garde, strong, cool, tough, strong,
Negative	Grumpy, dross, too slow, false, stacked, very annoying, worthless, loss, pit, weak mind, dissatisfaction, uneven, not straight, scarce, lacking, too bad, not necessary, poor, cheap, card, failure,

The emotional tendency of dynamic emotional words will change according to the different attributes and feature words it modifies. The same emotional word may express different emotional polarities in different contexts. For example, "big" is a negative word when describing noise, but a positive word when describing capacity. Based on the analysis of the car review text, combined with the expertise in the field of new energy vehicles, this paper constructs a dynamic sentiment dictionary and its collocated attribute words, as shown in Table 3.

Table 3 Examples of dynamic sentiment dictionaries

Attribute word-sentiment word tuples	Sentiment orientation
Speeding up, fast	1
Power consumption, fast	-1
Cost effective, high	1
Insurance, high	-1
Storage space, large	1
Abnormal noise, large	-1

Finally, leveraging the domain-specific sentiment dictionary, the attributes and associated sentiments within the reviews were systematically extracted. The average sentiment score for each attribute category of the product was quantified using the specified formula, and the resulting sentiment analysis outcomes are presented in Table

4.

Table 4 Review sentiment analysis results

Review	Product Attribute				
	X_1^{neg}	X_1^{pos}	...	X_n^{neg}	X_n^{pos}
$i = 1$	0	1		0	0
$i = 2$	1	0		0	1
...					
$i = m$	0	1		1	0

4.4 Customer demand identification

Determination of thresholds λ and δ . Following Qi et al. (2016), we initially set $\lambda = 0.1$ and $\delta = 0.1$. To assess the robustness of the Kano classifications, we conducted a sensitivity analysis by varying λ from 0.05 to 0.20 (step 0.05) and δ from 0.05 to 0.20 (step 0.05), yielding 16 parameter combinations. For each combination, we recomputed the Kano categories for the nine attributes. The results (summarized in Supplementary Table S1) show that eight attributes remained in the same category across all combinations. Only “price” switched between attractive and one-dimensional when $\lambda > 0.15$ or $\delta < 0.05$. Given that β_{price}^{neg} is not statistically significant ($p = 0.343$, see Table 5), the attractive classification is more conservative and theoretically sound. Therefore, we report the results with $\lambda = \delta = 0.1$ as the main findings, noting that the overall conclusions are robust to parameter variations.

Final classification. Based on the estimated coefficients (Table 5) and the decision tree with $\lambda = \delta = 0.1$, the product attributes were classified as shown in Table 6.

Table 5 Results of parameter estimation β_j

Variable	Estimate	Std. error	p value	Variable	Estimate	Std. error	p value
X_1^{pos}	0.073	0.016	0.000	X_5^{neg}	-0.013	0.024	0.000
X_1^{neg}	-0.012	0.033	0.025	X_6^{pos}	0.079	0.015	0.000
X_2^{pos}	0.060	0.023	0.005	X_6^{neg}	-0.040	0.042	0.343
X_2^{neg}	-0.020	0.030	0.028	X_7^{pos}	-0.003	0.019	0.894
X_3^{pos}	0.040	0.017	0.045	X_7^{neg}	-0.050	0.020	0.000
X_3^{neg}	-0.045	0.029	0.016	X_8^{pos}	0.064	0.015	0.000

X_4^{pos}	0.071	0.034	0.036	X_8^{neg}	-0.032	0.019	0.096
X_4^{neg}	-0.114	0.023	0.000	X_9^{pos}	0.058	0.015	0.000
X_5^{pos}	0.088	0.013	0.000	X_9^{neg}	0.034	0.020	0.090

Table 6 Product attributes for Kano classification

Product attribute	Classification Results
Driving range	Attractive Attribute
Space	Attractive Attribute
Appearance	One-dimensional Attribute
Intelligence	Attractive Attribute
Interior	Attractive Attribute
Price	Attractive Attribute
Seats	Must-be Attribute
Chassis system	Must-be Attribute
Driving experience	Must-be Attribute

In this study, nine key product attributes were identified and categorized according to the Kano model through thematic modeling analysis of new energy vehicle (NEV) review data. The findings indicate that these attributes vary in terms of their importance and influence on user experience.

Driving range, space, intelligence, interior, and price are classified as attractive attributes. Enhancing these attributes can substantially increase user satisfaction, while their underperformance has a relatively minor effect on user satisfaction. This suggests that these attributes represent areas where NEV brands can prioritize innovation and optimization to gain a competitive edge, attract more users, and enhance brand image. Appearance is categorized as a one-dimensional attribute, implying a direct correlation between appearance quality and user satisfaction. High-quality appearance design can elevate user satisfaction, whereas poor design may lead to dissatisfaction. Thus, appearance design is a critical factor influencing users' decisions regarding NEV

selection, and manufacturers should allocate significant resources to meet aesthetic expectations. Seats, chassis system and driving experience are classified as Must-be Attribute. These attributes reflect users' fundamental expectations for NEVs. Poor performance in these areas can significantly decrease user satisfaction, while even excellent performance struggles to markedly improve satisfaction. This highlights that these attributes serve as prerequisites for NEVs and constitute basic considerations for users when evaluating products. Manufacturers must ensure compliance with industry standards for these attributes to prevent losing users due to deficiencies in basic functionalities.

This study offers new energy vehicle (NEV) manufacturers guidance for optimizing product attributes based on user feedback. Manufacturers should prioritize enhancing attractive attributes, such as endurance, space, intelligence, interior, and price, to strengthen product competitiveness. Simultaneously, it is essential to ensure that necessary attributes, including seating arrangements, vehicle systems, and driving experience, meet fundamental standards to align with user expectations. Furthermore, appearance design, classified as a one-dimensional attribute, requires continuous improvement to increase the product's appeal. By implementing these strategies, NEV manufacturers can effectively address user needs, enhance user satisfaction, and gain a competitive edge in the highly contested market.

Following the classification of each car attribute, we evaluate the priority of improving specific car attributes by analyzing both importance and satisfaction levels. Importance is determined by the frequency of mentions for each attribute within the dataset, while satisfaction is assessed through sentiment scores associated with each attribute. Table 7 presents the chance scores for each car attribute.

Table 7 Chance score of each car attribute

Topics	Importance degree	Satisfaction degree	Opportunity
Space	0.88041766	0.827199494	1.70761621
Driving range	0.672940986	0.760577453	1.43351844
Driving experience	0.752631579	0.590953	1.34358453
Interior	0.685563679	0.568669546	1.25423323

Chassis system	0.689018948	0.344547886	1.03455674
Price	0.504172557	0.527199494	1.03137205
Seats	0.791852809	0.185035404	0.976888213
Appearance	0.38041766	0.614529829	0.75594749
Intelligence	0.252631579	0.252631579	0.505263158

From this table, it can be seen that the most prioritized improvements are space, endurance and driving experience, space and endurance are attractive attributes for users, and customer satisfaction will be greatly improved if improved. The essential attributes of driving experience are the most concerned attributes of users, however, there is still a lot of room for improvement according to the sentiment score.

5. Discussion

This study applied the proposed framework of dynamic sentiment lexicon construction integrated with the Kano model to analyze 8,970 online reviews of NEVs, and obtained clear results of customer needs mining, attribute classification and improvement prioritization. The Kano classification results show that driving range, space, intelligence, interior and price are attractive attributes, which can significantly improve user satisfaction once optimized, and are the core competitive points of NEV products. Appearance is a one-dimensional attribute, indicating that it is directly proportional to user satisfaction, which is an important factor for users' purchasing decisions. Seats, chassis system and driving experience are must-be attributes, which are the basic requirements of users, and insufficient performance will directly lead to user dissatisfaction.

The opportunity algorithm ranking results show that space, driving range and driving experience are the top three priorities for improvement. Among them, space and driving range belong to attractive attributes with high user attention and great optimization potential, while driving experience, as a must-be attribute, has low user satisfaction and needs to be focused on.

Compared with traditional static sentiment lexicons, the dynamic domain sentiment lexicon constructed by TF-IDF-PageRank and BERT contextual embedding

can better capture the context-sensitive emotional expressions in UGC, which effectively improves the accuracy of sentiment analysis. Compared with the traditional Kano model based on manual questionnaires, this data-driven method avoids subjectivity and small sample bias, and realizes automatic and efficient customer needs analysis. However, this study also has limitations: first, the research is only applied to the NEV field, and the cross-domain generalization effect needs to be further verified; second, fake reviews and low-quality reviews are not specifically identified and processed, which may have a slight impact on the results; third, only single-platform data is used, and multi-platform data fusion can be considered in the future to make the results more comprehensive.

In terms of practical value, this framework can help NEV manufacturers quickly capture user needs from online reviews, clarify product optimization priorities, reduce the cost of demand research, and provide a data basis for product iterative design.

6.Conclusion

This study proposes a complete customer needs analysis framework integrating dynamic sentiment lexicon construction and Kano model, which realizes the whole process of product attribute extraction, dynamic sentiment analysis, Kano classification and product improvement priority ranking based on online review data. The core conclusions are as follows:

First, a dynamic domain sentiment lexicon construction method based on TF-IDF-PageRank seed selection and BERT dynamic semantic representation is proposed, which solves the problem that traditional static lexicons are difficult to adapt to context-sensitive emotional expressions in UGC. Second, an end-to-end closed-loop framework for customer needs analysis is constructed, which realizes the objective and automated classification of customer needs by integrating dynamic sentiment analysis and Kano model. Third, the opportunity algorithm is used to realize data-driven product improvement priority ranking, which transforms user feedback into actionable optimization suggestions.

Validated on NEV online review data, the framework successfully extracts nine core product attributes, completes accurate Kano classification and improvement priority ranking, and has good practical application value. The limitations of this study include single-domain application, lack of fake review processing and single-platform data. Future research will expand the framework to cross-domain scenarios, integrate fake review detection models, and fuse multi-platform review data to further improve the stability and generalization of the method.

Acknowledgement

We are indebted to the anonymous reviewers and editor.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 72301035) and Beijing Municipal Commission of Education Science and Technology Plan Project (grant number KM202411232007).

References

- [1] H. Zhang, H. Rao, J. Feng, Product innovation based on online review data mining: A case study of huawei phones, *Electron Commer Res* 18 (2018) 3–22. <https://doi.org/10.1007/s10660-017-9279-2>.
- [2] R. Chen, Q. Wang, W. Xu, Mining user requirements to facilitate mobile app quality upgrades with big data, *Electronic Commerce Research and Applications* 38 (2019) 100889. <https://doi.org/10.1016/j.elerap.2019.100889>.
- [3] A. Naujoks, M. Benkenstein, Who is behind the message? The power of expert reviews on eWOM platforms, *Electronic Commerce Research and Applications* 44 (2020) 101015. <https://doi.org/10.1016/j.elerap.2020.101015>.
- [4] J.-W. Bi, Y. Liu, Z.-P. Fan, J. Zhang, Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews, *Tourism Management* 70 (2019) 460–478. <https://doi.org/10.1016/j.tourman.2018.09.010>.
- [5] S. Park, J.-S. Lee, J.L. Nicolau, Understanding the dynamics of the quality of airline service attributes: Satisfiers and dissatisfiers, *Tourism Management* 81 (2020) 104163. <https://doi.org/10.1016/j.tourman.2020.104163>.
- [6] J. Büschken, Sentence-based text analysis for customer reviews, (2016). <https://doi.org/10.1287/mksc.2016.0993>.
- [7] H. Liu, S. Wu, C. Zhong, Y. Liu, The effects of customer online reviews on sales performance: The role of mobile phone's quality characteristics, *Electronic Commerce Research and*

- Applications 57 (2023) 101229. <https://doi.org/10.1016/j.elerap.2022.101229>.
- [8] L. Zheng, Z. He, S. He, An integrated probabilistic graphic model and FMEA approach to identify product defects from social media data, *Expert Systems with Applications* 178 (2021) 115030. <https://doi.org/10.1016/j.eswa.2021.115030>.
- [9] L. Zheng, Z. He, S. He, Detecting and prioritizing product defects using social media data and the two-phased QFD method, *Computers & Industrial Engineering* 177 (2023) 109031. <https://doi.org/10.1016/j.cie.2023.109031>.
- [10] Z. Zhang, T. Luo, Product Demand Analysis Based on Online Review Data Mining and Kano Model, *Management Review* 34 (2022) 109-117. http://123.57.61.11/jweb_gipl/CN/
- [11] C. Berger, R. Blauth, D. Boger, C. Bolster, G. Burchill, W. DuMouchel, F. Pouliot, R. Richter, A. Rubinoff, D. Shen, M. Timko, D. Walden, Kano's methods for understanding customer-defined quality, in: 1993. <https://www.semanticscholar.org/paper/KANO%E2%80%99S-METHODS-FOR-UNDERSTANDING-CUSTOMER-DEFINED-Berger-Blauth/4fe74e5d7220a50662b272330e2f9ffc4e46d3ee> (accessed August 13, 2025).
- [12] Q. Meng, N. Zou, X. Chen, An" individual agent "evaluation method and its application in standard target system, *Management Review* 21 (2009) 86-93. <https://doi.org/CNKI:SUN:ZWGD.0.2009-12-013>
- [13] D. Baier, A. Rese, How to increase multichannel shopping satisfaction? An adapted kano based stage-gate approach to select new technologies, *Journal of Retailing and Consumer Services* 56 (2020) 102172. <https://doi.org/10.1016/j.jretconser.2020.102172>.
- [14] M.-C. Chen, C.-L. Hsu, C.-H. Huang, Applying the kano model to investigate the quality of transportation services at mega events, *Journal of Retailing and Consumer Services* 60 (2021) 102442. <https://doi.org/10.1016/j.jretconser.2021.102442>.
- [15] K. Sinemus, S. Zielke, T. Dobbelstein, Improving consumer satisfaction through shopping app features: A kano-based approach, *Journal of Retailing and Consumer Services* 85 (2025) 104243. <https://doi.org/10.1016/j.jretconser.2025.104243>.
- [16] H. Yu, C. Si, H. Wang, Research on product iteration and service innovation based on fuzzy Kano model, *Science and Technology Management Research* 38 (2018) 24-31. <https://doi.org/10.3969/j.issn.1000-7695.2018.09.004>
- [17] J.-W. Bi, Y. Liu, Z.-P. Fan, E. Cambria, Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model, *International Journal of Production Research* 57 (2019) 7068–7088. <https://doi.org/10.1080/00207543.2019.1574989>.
- [18] Y. Zhao, P. Ruan, X. Liu, X. Shan, Study on User Satisfaction Evaluation Based on Online Comment, *Management Review* 32 (2020) 179-189. <https://doi.org/10.14120/j.cnki.cn11-5057/f.2020.03.018>
- [19] M. Isnani, G.N. Elwirehardja, B. Pardamean, Sentiment analysis for TikTok review using VADER sentiment and SVM model, *Procedia Computer Science* 227 (2023) 168–175. <https://doi.org/10.1016/j.procs.2023.10.514>.
- [20] Z. Li, R. Li, G. Jin, Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary, *IEEE Access* 8 (2020) 75073–75084. <https://doi.org/10.1109/ACCESS.2020.2986582>.
- [21] N.R. Bhowmik, M. Arifuzzaman, M.R.H. Mondal, Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms, *Array* 13 (2022) 100123. <https://doi.org/10.1016/j.array.2021.100123>.

- [22] M. Zhang, L. Sun, G.A. Wang, Y. Li, S. He, Using neutral sentiment reviews to improve customer requirement identification and product design strategies, *International Journal of Production Economics* 254 (2022) 108641. <https://doi.org/10.1016/j.ijpe.2022.108641>.
- [23] C. Iwendi, S. Mohan, S. Khan, E. Ibeke, A. Ahmadian, T. Ciano, Covid-19 fake news sentiment analysis, *Computers and Electrical Engineering* 101 (2022) 107967. <https://doi.org/10.1016/j.compeleceng.2022.107967>.
- [24] C. Li, H. Liu, Association Analysis and N-Gram Based Detection of Incorrect Arguments, *Journal of Software* 29 (2018) 2243-2257. <https://doi.org/10.13328/j.cnki.jos.005531>
- [25] M. Arbane, R. Benlamri, Y. Brik, A.D. Alahmar, Social media-based COVID-19 sentiment classification model using bi-LSTM, *Expert Systems with Applications* 212 (2023) 118710. <https://doi.org/10.1016/j.eswa.2022.118710>.
- [26] B. Guo, C. Zhang, J. Liu, X. Ma, Improving text classification with weighted word embeddings via a multi-channel TextCNN model, *Neurocomputing* 363 (2019) 366–374. <https://doi.org/10.1016/j.neucom.2019.07.052>.
- [27] L.-W. Ku, Mining opinions from the web: Beyond relevance retrieval, *Journal of the American Society for Information Science and Technology* (2007). <https://doi.org/10.1002/ASI.20630>.
- [28] J. Chen, H. Lin, Z. Yang, Automatic acquisition of emotional vocabulary based on syntax, *Caai Transactions on Intelligent Systems* 4 (2009) 100-106. <https://api.semanticscholar.org/CorpusID:63877613>
- [29] Y. Zhao, B. Qin, T. Liu, Sentiment analysis: Sentiment analysis, *Journal of Software* 21 (2010) 1834–1848. <https://doi.org/10.3724/SP.J.1001.2010.03832>.
- [30] Y. Zheng, X. Jian, X. Zhuo, Utilization of sentiment analysis and visualization in online video bullet-screen comments, *Data Analysis and Knowledge Discovery* 31 (2015) 82 -90. <https://doi.org/10.11925/infotech.1003-3513.2015.11.12>
- [31] P.D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2001: p. 417. <https://doi.org/10.3115/1073083.1073153>.
- [32] Z. Wang, Z. Wu, F. Hu, Words Sentiment Polarity Calculation Based on HowNet and PMI, *Computer Engineering* 38 (2012) 187-189+193. <https://doi.org/10.3969/j.issn.1000-3428.2012.15.052>
- [33] B. Yang, Y. Liu, Y. Liang, M. Tang, Exploiting user experience from online customer reviews for product design, *International Journal of Information Management* 46 (2019) 173–186. <https://doi.org/10.1016/j.ijinfomgt.2018.12.006>.
- [34] N. Kano, N. Seraku, F. Takahashi, S. Tsuji, Attractive quality and must-be quality, *Journal of The Japanese Society for Quality Control* 14 (1984) 147–156. https://doi.org/10.20684/quality.14.2_147.
- [35] J. Qi, Z. Zhang, S. Jeon, Y. Zhou, Mining customer requirements from online reviews: A product improvement perspective, *Information & Management* 53 (2016) 951–963. <https://doi.org/10.1016/j.im.2016.06.002>.