International Journal of Industrial Engineering, 32(1), 237-252 , 2025

A CUSTOMER DEMAND MINING ALGORITHM BASED ON ONLINE COMMENTS AND MACHINE LEARNING

Yong Yang and Qiaoxing Li*

School of Management,
Guizhou University,
Guiyang, China.
*Corresponding author's e-mail: liqiaoxing23333@163.com

In the current market environment, the phenomenon of product homogenization is severe. If enterprises cannot deeply understand customer needs and provide differentiated products or services, it is difficult to stand out in the competition. In order to effectively improve overall customer satisfaction and enhance the market competitiveness of enterprises, a customer demand mining algorithm based on online comments and machine learning is proposed. Collect customer demand information data through online comments and process the collected data with redundant information to improve the efficiency and accuracy of demand mining. On this basis, customer demand attribute features have been further extracted, and a customer demand clustering mining model has been constructed using a self-organizing mapping neural network. By training the model, the final clustering mining results can be obtained, thus achieving precise mining of customer needs. This study clearly addresses a key issue in the current field of consumer demand mining: how to efficiently and accurately identify and utilize consumer demand information in online comments. By constructing a clustering mining model based on the Self-Organizing Maps (SOM) neural network, this study fills the literature gap in this field and provides more accurate and practical consumer demand analysis methods for enterprises. The experimental results show that, compared with the three comparison methods, the proposed method has a 98% feasibility of customer demand mining and 92% customer satisfaction. It shows that the proposed method has high feasibility and customer satisfaction for customer demand mining and has a better overall customer demand mining effect. This provides strong support for improving overall customer satisfaction and corporate competitiveness.

Keywords: Online Comments; Machine Learning; Customer Demand Mining; SOM Neural Network; Clustering Mining Model; Redundant Information Processing

(*Received on April 6, 2023; Accepted on August 24, 2024*)

1. INTRODUCTION

The rapid development of society has gradually increased customers' demands for personalized consumption. In order to gain competitive advantages, various industries have placed users' personalized needs in a very important position (Dai *et al.*, 2021). Many enterprises spend much effort on marketing strategies, marketing models, and other aspects because competition in the market ultimately comes from customers, and the existence of customers determines the production and sales of the enterprise. Under this promotion, the production and sales of enterprises can proceed in an orderly manner (Durowoju *et al.*, 2020). Nowadays, in order to meet the needs of customers, enterprises create greater customer value for customers than competitive enterprises so that customers can support their own enterprises and win the competition. According to the above analysis, it can be seen that if you want to obtain customer support, you must meet customer needs so that enterprises can survive and develop in the market competition. Therefore, it can be seen from the current research that the mining of customer demand (Wang and Zhou, 2021; Guney *et al.*, 2020) occupies an important position in enterprise product development, marketing, technical strategy, service support, organizational design, etc., and the development of enterprises is also related to product development and marketing. Therefore, in view of this description, in order to further promote the development of enterprises in the market environment, further mining and research of customer demand is required.

In order to solve the problems in the above methods and improve the market competitiveness of enterprises, research on customer demand mining algorithms based on online comments and machine learning is proposed. Using online comments as a data source for customer needs, these online comments directly reflect consumers' real feedback and expectations of products or services and have high real-time and directness. This data source not only enriches the dimensions of demand mining but also enables enterprises to respond more quickly to market changes and improve market sensitivity. On this basis, the redundant information in the collected data is eliminated, and the efficiency of demand mining is effectively improved

through redundant information processing. Extracting customer demand features from preprocessed data is a key step in understanding the true needs of customers. By constructing a clustering mining model, it is possible to group customer groups with similar needs, helping enterprises identify different segmented markets and providing strong support for customized marketing strategies and product innovation. Therefore, the introduction of neural networks and SOM (self-organizing map) clustering mining models, based on the powerful potential of machine learning in customer demand mining, can automatically learn complex patterns in data, achieve high-precision customer demand analysis and prediction, and provide a scientific basis for enterprise decision-making.

2. RELATED WORK

Nguyen *et al*. (2020) proposed a data mining method to predict customer demand for remanufactured products. The method provides a highly accurate and robust demand forecasting model for remanufactured products and elucidates the nonlinear effects of online market factors as predictors of customer demand. The demand for recycled products is predicted with high accuracy by using machine learning techniques, but customer satisfaction is lower as a result. Wang *et al*. (2022) proposed research on the demand mining of new energy vehicle consumers based on the fusion theme model. This research extracted the demand preference theme words of new energy vehicle consumers with the help of the theme model, calculated the similarity between the word vector and the theme words, expanded the theme words, analyzed and compared the demand themes and feature extension words of different models, and summarized the demand differences of other consumer groups. The analysis results show that this method can help consumers filter out valuable information from online comment data, help automobile companies objectively and accurately obtain consumer needs, formulate more reasonable marketing strategies, and achieve healthy and sustainable enterprise development. However, this method does not consider redundant information processing of the collected customer demand information data, resulting in low efficiency of customer demand mining. Peng *et al*. (2019) proposed a method for mining the demand of Chinese tourists for online short-term accommodation. This paper was devoted to exploring the demand for Chinese tenants in New York, Paris, Beijing, Shanghai and Guangzhou. This paper used Python language to grab about 130,000 comments on Airbnb as the basic data for this study, and extracted comments from Chinese guests. Based on the selection results, a two-dimensional analysis dimension was designed, and the feature words were classified into tables. Finally, the demand map was drawn according to the two-dimensional demand table. Through research, it is found that the needs of Chinese tenants can be divided into three aspects: internal environment, external environment and cultural needs. It enables foreign landlords to provide a better living environment for Chinese tenants, and at the same time, it can obtain more benefits and also enable Chinese tenants to have a better living experience abroad. However, this method does not extract customer demand attribute characteristics, resulting in low feasibility of customer demand mining. Liu *et al*. (2020) a method of product customer demand mining based on big data is proposed. The Hadoop platform is used to segment product attribute data, and feature word extraction based on the Apriori algorithm is used to mine product customer demand information. Apply the MapReduce model on the big data platform to efficient parallel data processing to obtain product attributes with research value and their weights and attribute levels. The cloud model and MNL model are used to build the product function attribute configuration model, the improved artificial bee colony algorithm is used to solve the model, and the optimal solution of the product function attribute configuration model is obtained. An example is used to illustrate the feasibility of the method in this paper, but the method does not consider the training of customer demand attribute characteristics, resulting in a poor overall mining effect of customer demand. Lin *et al*. (2020) a data mining algorithm of port customer loyalty based on Origin Destination (OD) data is proposed. Using the OD data, clustering is performed based on arrival date and port. The FP-growth algorithm is applied to mine frequent patterns of ships arriving at ports. Based on the frequent pattern and arrival time pattern of ships, the formula proposed in this paper is applied to push the shipping company's loyalty to the port. This method has certain effectiveness but still has the problem of low mining efficiency. Rozanec *et al*. (2021) created a recommendation system to help users make decisions and utilized development mechanisms to drive artificial intelligence. They enriched the knowledge map based on feedback modules and provided value to enterprises by accurately predicting and understanding the reasons behind the prediction, increasing confidence and helping decision-making. Zare *et al*. (2020) created a new concept based on real life and digital space called "digital marketing". By constructing the relationship between customers and management organizations, they established a data classification model and, based on this, conducted data mining to predict the actual needs of users. Arif and Hossain *et al*. (2021) used two text vectorization techniques, Bag of Words and TF-IDF, to convert text data into numerical feature vectors for real-time processing by machine learning algorithms, using user tweets and comments on Twitter as data sources. Train a classification model using a Support Vector Machine and Random Forest to mine customer feedback. Moazzam *et al*. (2021) collected customer data from e-commerce websites using web crawler technology and employed a combination of qualitative and quantitative e-commerce content analysis methods to conduct an in-depth analysis of the collected customer comments. Based on Bag of Words and N-Gram technology, features are extracted

from text data, and customer comments are classified using a naive Bayesian classifier to mine customer opinions. Neither of these methods handles redundant information, which can reduce the credibility of the results.

3. CUSTOMER DEMAND DATA COLLECTION, PROCESSING AND FEATURE EXTRACTION BASED ON ONLINE COMMENTS

Using online comment methods to collect customer demand data and eliminate redundant information, a word vector model based on Huffman Trees (HS) is constructed. The word vector method is used to extract customer demand attribute features, achieving efficient data collection.

3.1 Customer demand data collection and preprocessing based on online comments

(1) Online comments

When the online review method (Zhang *et al*., 2020) collects customer demand data (Liu *et al.*, 2021), it needs to be divided into several different elements, namely:

1. The virtual community can be used to obtain online information sharing among customers;
2. The customers' subjective perception opinions on enterprise products in the public network;
3. Communication between customers on the Internet about the specific characteristics, use effects and proposed conditions of enterprise products or services;
4. Evaluation of information generated by customers on product and service quality in e-commerce websites;
5. Online content generated spontaneously by products purchased by customers from e-commerce or third-party platforms and websites;
6. According to the current way of spreading word of mouth on the Internet, potential customers will be an important source of product information.

According to the above six elements, customer demand data based on online comments can be collected.

(2) Customer demand data collection

Among the global websites, the Meituan website is one of the most influential group-buying websites in China. Nearly hundreds of millions of customers browse the website interface every day, and the website service module is an important platform for customers to communicate and share with each other. In the website service module, a large number of users comment on related products every day. Therefore, the Meituan website is taken as the online comment data source of customer needs (Kauffmann *et al.*, 2020; Traub *et al.*, 2021), and the crawler program is used to capture the comment information on the website and obtain the effective data of customer needs.

Based on the review data of the Meituan website in Qingdao, food stores, hotels, fruit and vegetable stores, etc., are selected as the research objects, and 180 enterprise stores are selected as the customer review collection objects, including 12,453 comments from 48 food stores, 13,548 comments from 92 hotels, and 2,413 comments from 40 fruit and vegetable stores.

(3) Redundancy processing of customer demand data

Based on the customer demand data collected by online reviews, there is a large amount of redundant information in the collected data center. Therefore, in order to effectively realize customer demand mining, it is necessary to conduct redundant processing on the collected customer demand data (Che *et al.*, 2021).

According to the customer demand data collected by online reviews, if enterprise researchers mine factual comments and opinion comments on online enterprise products, they will ignore the fact that the data mining information already exists in the database. In this way, the extracted customer demand information will lose its value and consume a lot of time in data mining (Yang and Feng, 2021). Therefore, after collecting the customer demand data, a reasonable method should be selected to eliminate the redundant data. Therefore, a redundant information processing model should be built based on personalized demand acquisition. The construction results are shown in Figure 1.

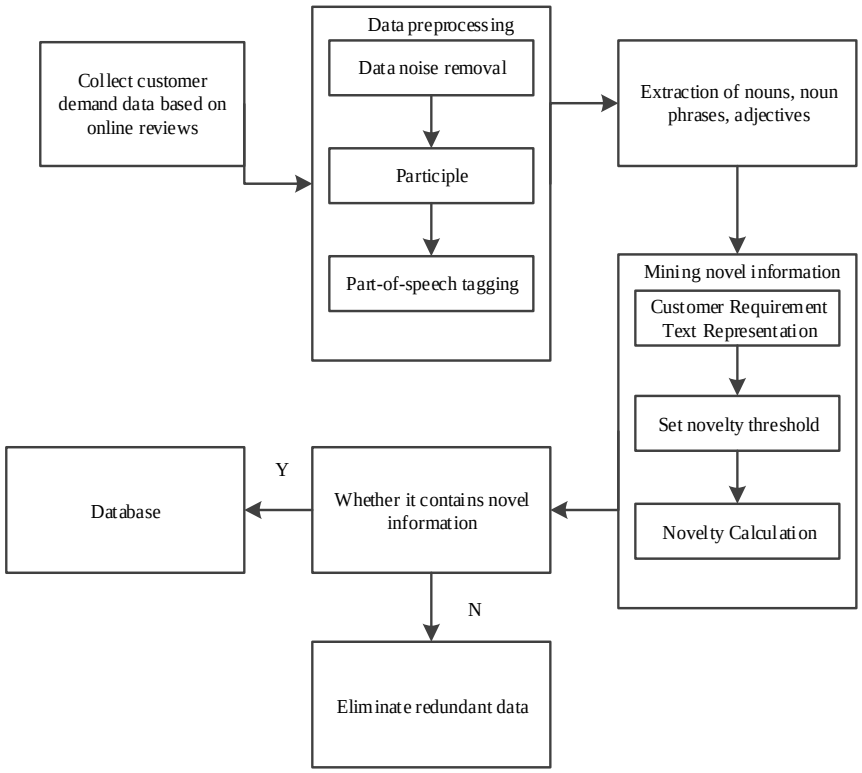

Figure 1. Redundant Information Processing Model Based on Personalized Demand Acquisition

According to the constructed model, the collected online customer demand data is arranged according to the order of time and is defined as: $(s_1, s_2, s_3, \cdots, s_n)$, where each customer demand review record is defined as $s_i (i = 1,2,3, \cdots, n)$. Where n is the number of articles and i is the comment coefficient.

The novelty value of each comment is measured according to the historical comment record sentences of the website, so it is necessary to set a novelty threshold α to judge the novelty of the collected comment sentences s_i of customer demand. When the novelty value of s_i is higher than the set α, it indicates that the obtained customer demand comment sentence is novel. At this time, it is necessary to store the comment record in the delivery history database; On the contrary, if the novelty value of s_i is lower than α, the comment record s_i is discarded, thereby eliminating redundant information in the customer demand comment data.

It is set that the best quality of the product is defined by e, the low price is defined by f, and the high-cost performance is defined by g. Therefore, there is a record in the database marked as $s_1 = (e, f, g)$. If the appearance of the product is defined by h, the volume is defined by i, and the high resolution is defined by j, then a record is re-acquired as $s_2 = (e, h, i, j)$. Since s_2 has three infrequently occurring words, the novelty of s_2 is set to 72%. Then, the setting result is compared with the previously set α, and it can be seen that when $\alpha = 0.5$, s_2 belongs to a novel word; When $\alpha = 0.7$, s_2 is not a novel word.

According to the above requirements, a vector space model is established, which is mainly used in text representation statistics and belongs to the statistical model. It can record online comments of customer needs and use them as text vectors, which are defined as:

$$s = [w_1(s), w_2(s), w_3(s), \cdots, w_n(s)]^T \tag{1}$$

In formula (1), $s$represents a text vector, which is also a comment record, $w_1(s)$represents a new record sub-vector, and T represents time.

In order to judge the novelty between the new record and the historical record, the cosine similarity method is used to judge the two kinds of records. The s_d represents the new comment record and $s_t$represents the historical comment record, and then the similarity between the two can be calculated. The calculation equation is as follows:

$$cos(s_d, s_t) = \frac{\sum_{k=1}^{n} w_k(s_d) + w_k(s_t)}{\|s_d\| \cdot \|s_t\|} \tag{2}$$

In formula (2), $w_k(s)$ is the feature weight of the comment record whose k attribute is s.

Feature weights are mainly used to test the proportion of customer demand opinions in the document. When the comment feature items appear more times in the document, the feature weight can be calculated. In short, it is also called word frequency.

Therefore, when the comment vocabulary has a great impact on a document, the number of occurrences of the k feature word in s_i can be calculated, and the calculation expression is as follows:

$$\begin{cases} w_k(s) = f_{ik} \cdot f_i \\ f_i = \dfrac{N}{n_i} \end{cases} \tag{3}$$

In formula (3), f_{ik} represents the number of comment records, f_i represents the quantification of feature word t_i in all comment records, N represents the number of all comment records, and n_i represents the number of feature word records included in all comment records.

Therefore, according to formula (3), it can be seen that the calculation equation of customer demand comment novelty is:

$$NC(s_d) = 1 - \max_{1 \le t \le d-1} (cos(s_d, s_t)) \tag{4}$$

In formula (4), $NC(s_d)$ represents novelty, and d represents the novelty recording coefficient.

Since the setting of the novelty threshold α of customer demand comment is different, the result of obtaining the novelty of customer demand comment is also different. Therefore, the $NC(s_d)$ to be obtained, and the α to be set are compared with each other. When the result of $NC(s_d)$ is greater than α, it indicates that the customer demand comment record at this time belongs to novel information and needs to be kept; On the contrary, when the result of $NC(s_d)$ is less than α, it indicates that the customer demand comment record has no available value and belongs to redundant information, which needs to be removed or discarded.

3.2 Extraction of the attribute characteristics of customer demand comment data

(1) Construction of word vector model based on HS

By eliminating the redundant information of customer demand comments, it can use the word vector method to extract the attribute features of customer demand (Alsenan *et al.*, 2020).

The word vector is set as $Context(w)$, the dimension of $Context(w)$ as M, and $Context(w)$ is mainly composed of c vectors before and after M. In the artificial neural network, the word vector model is often used to train the collected customer demand comment data and extract the attribute features of the customer demand comment data. Therefore, when $Context(w)$ is in the input layer of the artificial neural network, its own context word vector is marked as: $V(Context(w)_1), V(Context(w)_2), \cdots, V(Context(w)_{2c})$. The projection layer is used to sum the $2c$ vectors in the input layer, and the average is obtained after accumulation. At this time, the output layer corresponds to the binary Huffman tree, and the word vector is formed after projection in the root node, and then the leaf node is used to train the word vector. Taking the words appearing in the corpus as leaf nodes, the number of leaf nodes is the size of the vocabulary. Therefore, when the number of leaf nodes of the binary Huffman tree is N, the number of non-leaf nodes is $N - 1$.

Through the basic description of the internal network of the training word vector model, the HS-based training word vector model (Zhang *et al.*, 2020), i.e. continuous bag of words (CBOW) model, is constructed. The construction results are shown in Figure 2.

According to the processed online comment data of customer demand, the customer demand training sample is obtained from the data and is marked as: $(Context(w), w)$. At this time, the objective function of the training sample data is defined as:

$$L = \prod_{w \in corpus} P\big(W | Context(w)\big) \tag{5}$$

In formula (5), L represents the objective function, W represents any comment word in the vocabulary, P represents the path, and $corpus$ represents the corpus.

From formula (5), it can be seen that W in the vocabulary must have a path corresponding to W from the root node in the Huffman tree, that is, p^w, and p^w is the only path in the Huffman tree. When there are $l^w - 1$ branches in path p^w, where

l^w represents the number of nodes in path p^w, each branch needs to be classified twice, and each classification will generate a probability. Therefore, when each node of the Huffman tree has a corresponding category after the two classifications, it is necessary to mark the left subtree as a negative category and set its code as 1. Conversely, it marks the right subtree as a positive class and sets its code to 0. Therefore, according to the setting conditions, the positive class or negative class is judged as follows:

$$\sigma(x_w^T\theta) = \frac{1}{1+e^{-x_w^T\theta}} \tag{6}$$

In formula (6), x_w^T represents the word vector in the current internal node, θ represents the model parameter calculated in the training sample, σ represents the probability, and e represents the number of nodes. Where, $\sigma(x_w^T\theta)$ represents the probability of judging the positive class, and $1-\sigma(x_w^T\theta)$ represents the probability of judging the negative class. These two parameters can be judged by the positive and negative probability values of a node to the left subtree and the right subtree.

Therefore, the root node is the main node. When performing the secondary classification on all non-leaf nodes of the word W in the path p^w, it needs to determine the probability accumulation product that needs to be selected to obtain $P(W|Context(w))$.

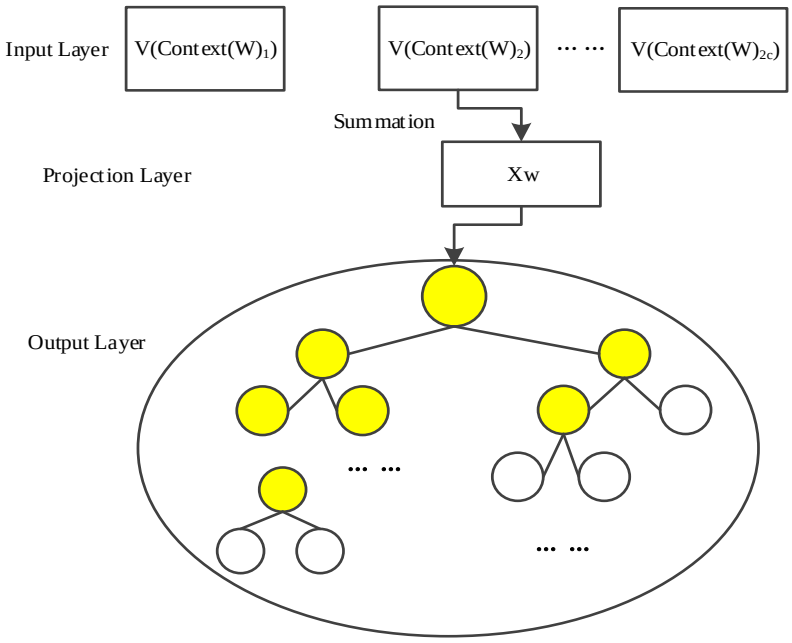

Figure 2. Constructed CBOW Model

The HS-based word vector model is mainly used to find the appropriate comment node vector and auxiliary vector of customer demand, and θ is used to maximize the objective function value to ensure the highest output probability of the activation function softmax. Therefore, the specific solution process is as follows:

1. Sum the $2c$ word vectors within the range of W, and obtain the average value from them, which is calculated as:

$$x_w = \frac{1}{2c}\sum_i^{2c} x_i \tag{7}$$

In formula (7),x_w represents the sum result of word vectors, and x_i represents the average value.

2. The random gradient rising method is used to iterate the auxiliary vectors θ_{j-1}^w and x_w, which are expressed as follows:

$$\begin{cases} \theta_{j-1}^{w} = \eta\left(1 - d_j^w - \sigma(x_w^T\theta)\right)x_w \\ x_w = \eta\sum_{j-1}^{w}\left(1 - d_j^w - \sigma(x_w^T\theta)\right)\theta_{j-1}^{w} \end{cases} \tag{8}$$

In formula (8), η represents the number of iterations.

According to the above two steps, the customer demand comment words can be mapped to form a word vector with a fixed length, and all these word vectors can be attributed to the same space to form a word vector space. In the space, each word vector is a point in the space, and the distance between each word vector reflects the correlation between customer demand comments. From the word vector distance calculated by the cosine distance, it can be seen that the closer the calculated cosine distance value is to 1, the greater the correlation between the comments is, and vice versa. Therefore, according to this feature, an internal command is set in the established word vector model, and the command is used to obtain the synonyms between comment words. Therefore, after the model is used to train the customer demand comment words, a comment word should be re-entered, and the list of comment words closest to the input words can be obtained.

(2) Extraction of the attribute characteristics of customer demand comments

According to the training word vector model constructed above, the attribute features of customer demand comment data are extracted by using the model. The extraction process is as follows:

1) After the customer demand data based on online comments is processed, the processed text comment data should be first analyzed;
2) Use the word vector model in gensim to train the text comment corpus after word segmentation;
3) Set the sliding window $window = 10$, the dimension of the word vector $vetor = 300$, and map all the text comments into the set 300-dimensional vector space to form the corresponding word vector;
4) Use the internal command of the word vector model to obtain the list of synonyms of customer demand comments, input a comment word into the model, and obtain the word list and cosine distance close to the input word;
5) Input the seed words of customer demand comments and form the attribute feature library of customer demand comments according to the similarity of the words;
6) Select the seed words of the attribute characteristics of customer demand comments for similarity matching (Kim *et al.*, 2019; Zhang *et al.*, 2020), and extract the related words of the seed feature words to realize the extraction of the attribute characteristics of customer demand comments.

4. CUSTOMER DEMAND MINING BASED ON MACHINE LEARNING

Customer demand indicators are established and their values are quantified based on the above extraction results. Using machine learning methods (Liu *et al.*, 2022) to construct a SOM clustering mining model, the extracted customer demand attribute features are input into the model for training, and clustering is performed based on customer needs to achieve high-precision mining of customer needs (He and Yin, 2021).

4.1 Selection and quantification of customer demand indicators

In order to further mine customer demand, customer demand indicators and quantification should be established.

Suppose an enterprise needs to sell a total of m products, defined as: $C = \{c_1, c_2, \cdots, c_m\}$, C represents all the products sold. Each product needs to have n different attributes, i.e., feature indicators, and the attribute set of each indicator is expressed as: $S = \{s_1, s_2, \cdots, s_n\}$. The value mark V_{ij} of the product c_i in the attribute s_i is defined as $V = \{V_{ij}\}m \times n$ by the expression.

When there are k customers $X = \{x_1, x_2, \cdots, x_k\}$ in the enterprise, the product set purchased by customers $x_1(1 = 1,2, \cdots, k)$ in period T is $C^1(C^1 \subseteq C)$. Generally, $C^1(C^1 \subseteq C)$ mainly includes product quality, product price, product function, product style and product fashion, so the attribute values of each product are different in different $C^1(C^1 \subseteq C)$. Therefore, it can be seen that similar products A and B have high and low quality. A good service level and product attributes not only save the marketing expenses of the enterprise but also enhance customer satisfaction.

In view of this feature, the customer demand characteristics and demand levels are divided. Therefore, four indicators, such as product cost performance, product style, product price and product brand are quantified, and the remaining indicators are qualitative indicators. When quantifying the indicators, the expert scoring method can be used to score and judge the products, and the average value can be selected as the quantitative value of the product in a certain indicator.

4.2 Establishment of clustering mining model based on machine learning

(1) Machine learning

As an interdisciplinary subject, machine learning (Akbar *et al.*, 2021; Kosasih and Brintrup, 2022) is to design or analyze some algorithms that can enable computers to learn automatically, and obtain certain laws from them, and then use this law to mine data. It can be said that machine learning is the most critical part of artificial intelligence.

In machine learning algorithms it is divided into supervised learning and unsupervised learning. Supervised learning includes support vector machine, artificial neural network, logistic regression, decision tree, K-nearest neighbor and other algorithms; Unsupervised learning includes K-means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering and principal component analysis.

(2) Constructing SOM clustering mining model

According to the characteristics of different algorithms, neural network clustering based on machine learning is selected for deep training and mining of customer demand features.

Self-organizing mapping neural network, i.e., SOM neural network, is a common clustering neural network (Zheng and Ma, 2021; Wang *et al.*, 2021). It is mainly composed of an unsupervised learning neural network model. Its function is to map the input N -dimensional spatial data to a lower dimensional output and maintain the original topological logical relationship of the data. According to the automatic organization of neural network neurons, the SOM network is used to map the extracted attribute characteristics of customer demand. According to the mapping, the distribution and classification of various types of data are obtained to complete the mining of customer demand.

There are two levels in the SOM network, i.e., the input layer and the output layer. When neurons are transported to the input layer, they are connected to each neuron of the output layer. Then, the clustering mining model based on the SOM neural network and machine learning is constructed, as shown in Figure 3.

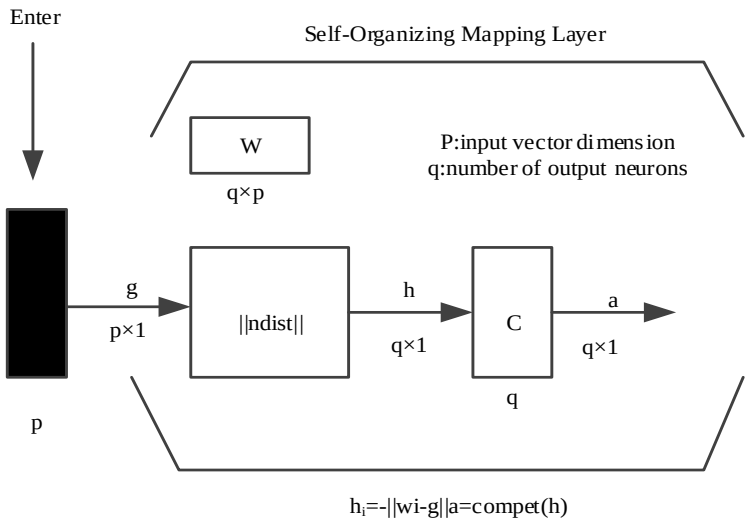

Figure 3 SOM Clustering Mining Model

It can be seen from Figure 3 that in the constructed SOM model, the neurons in the input layer are mainly arranged in one-dimensional form, the number of input neurons is determined by the number of components in the input vector, and the neurons in the output layer are arranged in one-dimensional or two-dimensional form, from which the number of neural elements in the input layer is calculated, expressed as p, and the number of neurons in the output layer is q, so that it meets the condition of $q >> p$.

When the total number of input samples in the SOM model is k, the vector expression equation of the first input sample in the model is defined as:

$$X^1 = \left(x_1^1, x_2^1, x_3^1, \cdots, x_p^1\right)^T \tag{9}$$

The output value of each output neuron is set as $y_j, j = 1,2,\cdots,q$, where j represents a coefficient. to obtain the weight vector between the output value of the output neuron and the j output neuron, and it is defined with the equation expression as follows:

$$W_j = \left(x_{j1}, x_{j2}, x_{j3}, \cdots, x_{jq}\right)^T \tag{10}$$

According to equation (9) and equation (10), the SOM clustering method is to find out each input neuron and its corresponding output neuron and establish a new neuron, i.e., the optimal neuron, by using the best matching method of input vector and weight vector. There is the nearest Euclidean distance between the optimal neuron and the input sample, so when z represents the winning unit of the first sample, the following conditions can be met, which is defined as follows:

$$\|X^1 - W_Z\| < \left\|X^1 - W_j\right\|, j \subset q, j \neq z \tag{11}$$

In formula (11), W_Z represents the first winning sample unit.

4.3 Clustering mining of customer demand based on SOM neural network and machine learning

According to the constructed clustering mining model based on SOM neural network, the customer demand index variables and input values are calculated, the extracted attribute features of customer demand are input into the model for training (Liu *et al.*, 2020; Wang *et al.*, 2020), and the customer demand clustering is realized according to the training results to complete the customer demand mining.

(1) Calculate index variable standardization and input value

There is incommensurability between different product indicators in an enterprise. When the price of class A product is low, and the price of class B product is high, the SOM neural network model can be used to mine data features at the same time. Therefore, based on the transfer function, the extracted data features are determined within the range of interval [0, 1]. When the index variables are standardized, they are also made within the range of the set interval [0, 1].

The maximum value of an attribute indicator of a product is set as A_{max}, the minimum value as A_{min}, the standardized value as B, and the initial value as C. Then, the attribute indicator value of the standardized product is calculated by the following equation:

$$B = \frac{C - A_{min}}{Amin_{max}} \tag{12}$$

Customers need to purchase different products in a certain period of time. For example, when customer x_1 purchases $m_1(m_1 < m)$ products, the standardized value of the price of each product is expressed as: $v_i(i = 1,2,\cdots,m_i)$. Then, the customer's cognitive value of product price can be obtained through the following equation, marked as:

$$V' = \frac{1}{m_1}\left(\sum_{i=1}^{m} V_i\right) \tag{13}$$

In formula (13), V' represents the cognitive value.

The 4-dimensional vector value corresponding to each customer can be calculated in this way as follows:

$$V_1 = (v_{11}', v_{12}', v_{13}', v_{14}') \tag{14}$$

In formula (14), v_{11}' represents the value of the customer's cognition of cost performance, v_{12}' represents the value of the customer's cognition of style, v_{13}' represents the value of the customer's cognition of price, and v_{14}' represents the value of the customer's cognition of brand.

According to the obtained results, the 4-dimensional vector value corresponding to the user and the extracted attribute feature of customer demand are used as the input of the clustering mining model based on the SOM neural network to realize the clustering of customer demand.

(2) Determine the structure and training of the neural network

In the SOM neural network, the number of neurons in the input layer is determined by the number of classification measurement indicators, so the number of neurons in the input layer is set to 5 according to the obtained 4 measurement indicators and the extracted customer demand attribute characteristics. The number of neurons in the output layer is determined to be $2 \times 2 \times 2 \times 2 \times 2 = 32$. Each index value is divided into two categories: greater than the average value and less than the average value, then the distribution of output neurons in the two-dimensional space is: $5 \times 5 = 25$.

Matlab software is used to construct the usable function form in the SOM neural network, and the definition is as follows:

$$net = newsom(PR, D, TFCN, DFCN, OLR, OSTEPS, TLR, TND) \tag{15}$$

In formula (15), net represents network, $newsom$ represents a function form, and PR represents a $P \times 2$-dimensional matrix composed of the possible minimum and maximum values of each dimension in the P -dimensional input amount; D represents the number of arrangement of neurons in the output layer in the multi-dimensional space, $TFCN$ represents the topological function, $DFCN$ represents the distance function, OLR represents the stage learning rate, $OSTEPS$ represents the number of learning times, TLR represents the adjustment learning rate, and TND represents the neighborhood radius of the adjustment stage.

In the SOM neural network clustering mining model, the function $learnsom$ is used as the model learning rule. Therefore, when the training function in the SOM neural network is $trainr$, or the adaptive function is set to $trains$, it is necessary to use the scheduling function $net = train(net, P)$ to train the index value in the neural network and the extracted customer demand characteristics. In expression $net = train(net, P)$, P represents an input vector matrix.

(3) Clustering mining process of customer demand

1. Use $(v_{11}^{'}, v_{12}^{'}, v_{13}^{'}, v_{14}^{'})$ as the data to divide customers, and use part of the feature data of customer demand as the training data set; Use the clustering ability of SOM neural network to cluster multiple customer demand clusters, that is, 16 clusters.
2. The average value of the index characteristics $(v_{11}', v_{12}', v_{13}', v_{14}')$ of each customer demand cluster is expressed as: $(v_{L1}', v_{L2}', v_{L3}', v_{L4}')$, where L represents the L -class customer, i.e., $L = 1,2,\cdots,16$, then calculate the total average value of the corresponding demand of all customers, and mark it as $(\bar{V}_1, \bar{V}_2, \bar{V}_3, \bar{V}_4)$.
3. Compare the average value of index characteristics and total average value of each customer demand cluster. When comparing the characteristic average value and the total average value of the indicators, there will be two results, which are greater than or equal to the total average value and less than the total average value. According to the two comparison results, the indicator changes of each customer demand cluster are obtained, and the demand of each customer is subdivided according to the changes.
4. Conduct detailed analysis on the nature of customer demand clusters from the obtained changes of indicators of each customer demand cluster, such as whether the analyzed customer demand cluster is an actual consumer or a consumer pursuing quality.
5. Use the clustering mining model based on the SOM neural network after training to classify all feature data of customer demand (Memis *et al.*, 2021) so that each customer belongs to one customer demand type.

Finally, according to the above clustering mining processes of customer demand, the research on customer demand mining based on online comments and machine learning is realized.

5. EXPERIMENT AND DISCUSSION

In order to verify the overall effectiveness of the research on customer demand mining algorithms based on online comments and machine learning, it is necessary to carry out experimental comparative tests on this method. Using the four-dimensional vector value V_1 as the experimental training dataset, the proposed method, the method of reference (Van Nguyen *et al.*, 2020), the method of reference (Wang *et al.*, 2022) and the method of reference (Peng *et al.*, 2019) are used for comparative experiments. The raw data contains customer evaluations or demand information for various products, which exist in the form of text comments, survey questionnaires, transaction records, etc. It contains issues such as noise, missing values, inconsistent formats, etc., and requires preprocessing before it can be used for analysis. Data preprocessing refers to a clean dataset of data that has been cleaned, transformed, and standardized, where the features have been quantified and prepared as inputs for machine learning models. The data preprocessing steps are as follows:

Step 1 : Determine the data source of online comments, such as e-commerce platforms, social media, corporate websites, etc., and use web crawling technology to capture customer comment data;
Step 2 : Check and delete duplicate or highly similar comments to avoid interfering with the analysis results, fill in missing values, and identify and correct noisy data in the comments;
Step 3 : Perform word segmentation, stop word removal, stem extraction, or morphological restoration on the comment text for subsequent feature extraction;
Step 4 : Convert text comments into high-dimensional vector representations to capture semantic relationships between words;
Step 5 : Select the features with higher correlation by calculating the correlation between the features and the target of demand mining;
Step 6 : Divide the preprocessed data into a training set and a testing set for subsequent training and evaluation.

In order to effectively test the effect of customer demand mining, it should randomly obtain the customer demand survey of an enterprise and obtain the demand information table of 11 customers for three types of attributes, including price a_1, weight a_2 and purpose a_3. Details are shown in Table 1.

Table 1. Intention of Customer Demand Information

Customer	a_1/ Ten thousand yuan	a_2/t	a_3
x_1	50	18	1
x_2	210	35	3
x_3	60	12	2
x_4	80	24	1
x_5	14	3	1
x_6	150	38	2
x_7	140	35	2
x_8	340	67	3
x_9	10	2	1
x_{10}	50	10	1
x_{11}	40	5	2

The usage values a_3 of the customer demand attribute are 1, 2 and 3. When $a_3 = 1$, it represents the daily demand of the customer; When $a_3 = 2$, it represents the daily consumption products of customers, while when $a_3 = 3$, it represents luxury goods.

According to the customer demand information obtained in Table 1, the proposed method, the method of reference (Van Nguyen *et al.*, 2020), the method of reference (Wang *et al.*, 2022) and the method of reference (Peng *et al.*, 2019) are used to carry out the feasibility comparative test of customer demand mining and customer satisfaction, and the performance of customer demand mining of the four methods is verified based on the test results.

(1) Feasibility comparison test of customer demand mining

To test the feasibility of customer demand mining, it is necessary to calculate the feasibility of customer demand mining by using the following equation expression, which is defined as follows:

$$F_\lambda = w_s S_\lambda + w_c C_\lambda \tag{16}$$

In formula (16), F_λ represents the feasibility index of customer demand mining, S_λ represents the average customer satisfaction for the threshold λ, C_λ represents the cost ideality of enterprise for the threshold λ, w_s represents the weight of customer satisfaction in the feasibility index and w_c represents the weight of the enterprise's cost ideality in the feasibility index, i.e., $w_s + w_c = 1$.

The specific test results are shown in Figure 4.

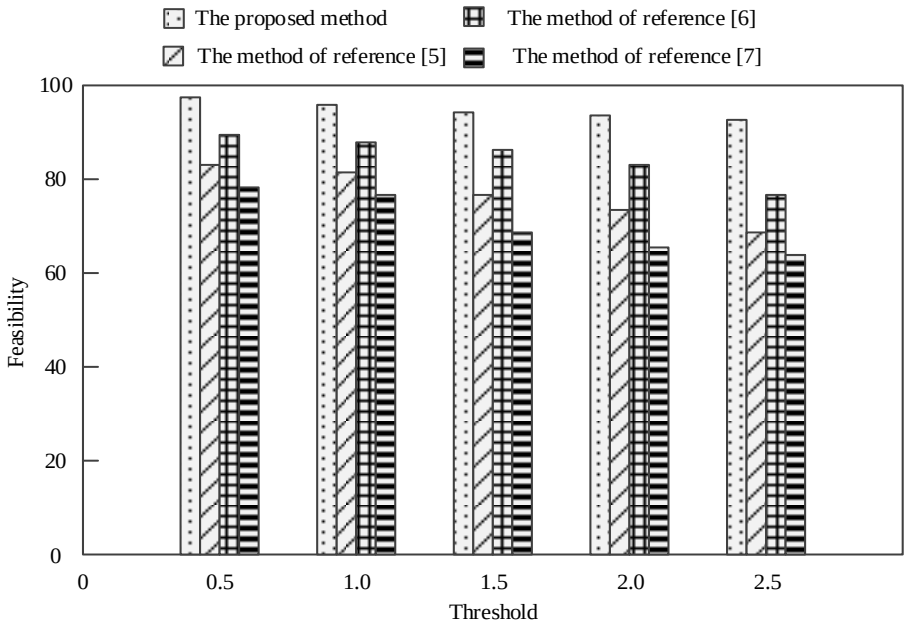

Figure 4. Feasibility Comparison Test of Customer Demand Mining

The number of clusters will continue to increase with the increase of the threshold value λ, and a high threshold value λ will make customer satisfaction higher and higher, but the feasibility index F_λ of customer demand mining will be lower and lower.

According to this feature, it can be seen from the test results that when the threshold value is small, the feasibility of the proposed method is higher. At the same time, as the threshold continues to increase, the feasibility indicators of the four methods show a downward trend. After comparing the four methods, the feasibility of the proposed method is the highest in the entire test, always above 90, and the maximum difference in feasibility indicators is more than 30. Therefore, it can be seen that the proposed method is more feasible for customer mining.

(2) Comparison test of customer satisfaction

In order to further test the satisfaction of customer demand mining, s_i is set as the satisfaction of the i customer to customer demand mining, then the satisfaction of the i customer is defined as:

$$s_i = \frac{\sum_{k=1}^{l}\left(1-\frac{h-x_{ik}^{'}}{h}\right)+\sum_{k=l+1}^{n}\frac{n_{ik}}{n_i}}{n_i} \tag{17}$$

In formula (17), h represents the customer satisfaction attribute, n_{ik} represents the number of customers with the same equivalent attribute in the class of the i customer, n_i represents the number of customers, k represents the attribute, $x_{ik}^{'}$ represents the i customer attribute, and l represents the coefficient.

Then, the calculation expression of customer satisfaction is defined as follows:

$$S_\lambda = \frac{\sum_{i=1}^{m} s_i}{m} \tag{18}$$

In formula (18), S_λ represents customer satisfaction, and m represents the attribute assignment.

The following tests are conducted on customer demand satisfaction by using the four methods.

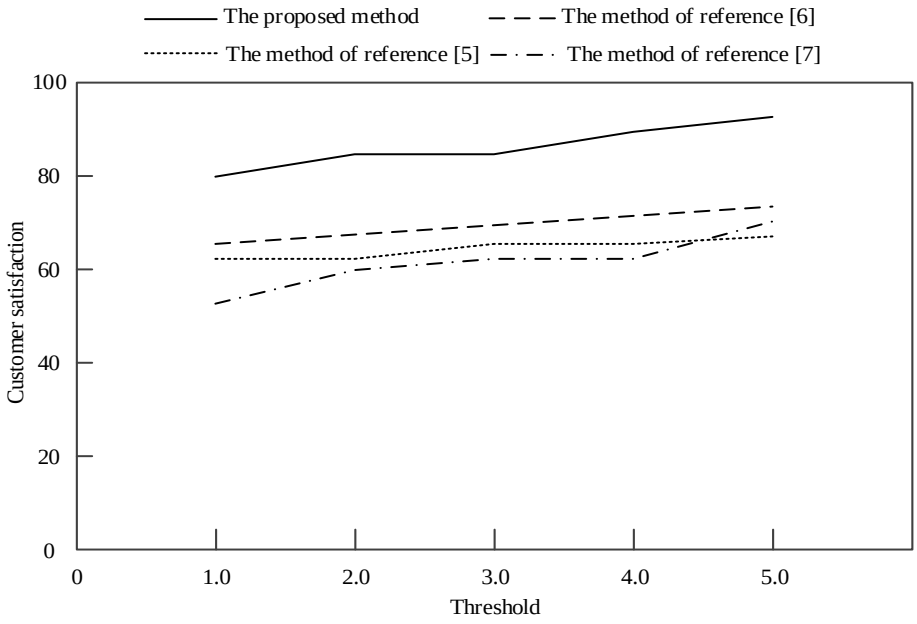

Figure 5 Comparison Test of Customer Satisfaction

From the data in Figure 5, as the threshold continues to increase, the customer satisfaction of the proposed method remains high, reaching a maximum of over 90 and a minimum of around 80. Comparing the satisfaction of the four methods under different thresholds, it can be found that the satisfaction of the proposed method is higher than that of the other three methods, indicating that the proposed method has a good overall customer demand mining effect.

The four methods are used to compare and test the running time of customer demand mining, as shown in Table 2.

Table 2. Comparative Testing of the Running Time of Customer Demand Mining by Four Methods

Customer demand information data volume/MB	The method of reference [5]/s	The method of reference [6]/s	The method of reference [7]/s	The proposed method/s
1000	4.3	5.6	3.8	1.4
2000	5.2	6.8	4.6	1.9
3000	6.9	7.5	5.9	2.3
4000	7.4	8.9	6.4	2.8
5000	8.3	9.7	7.8	3.4

According to Table 2, as the amount of customer demand information data increases, the running time of different methods of customer demand mining increases. When the data volume of customer demand information is 5,000MB, the running time of the method of reference (Van Nguyen *et al.*, 2020), the method of reference (Wang *et al.*, 2022) and the method of reference (Peng *et al.*, 2019) are 8.3s, 9.7s and 7.8s respectively. The running time of the proposed method for customer demand mining is only 3.4s. It can be seen that the customer-demand mining running time of the proposed method is shorter, and its customer-demand mining efficiency is higher.

On this basis, a formal statistical analysis is performed on the experimental results, and the p-value is set to 0.05. The smaller the p-value, the more significant the results are. The p-value of the experimental results of the proposed method is 0.02, and the p-value is less than 0.05, and the significance level is statistically significant.

To sum up, the feasibility and satisfaction of customer demand mining of the proposed method are better than those of the method of reference (Van Nguyen *et al.*, 2020), the method of reference (Wang *et al.*, 2022) and the method of reference (Peng *et al.*, 2019). The efficiency of customer demand mining is improved, and its feasibility and satisfaction are further enhanced due to the proposed method's processing of redundant information in the collected customer demand comment data.

In order to further test the accuracy of customer demand data mining for the proposed method, the F1 values of four methods are tested. The range of F1 values is between [0,1], and the larger the value, the higher the data mining accuracy. The test results are shown in Figure 6.

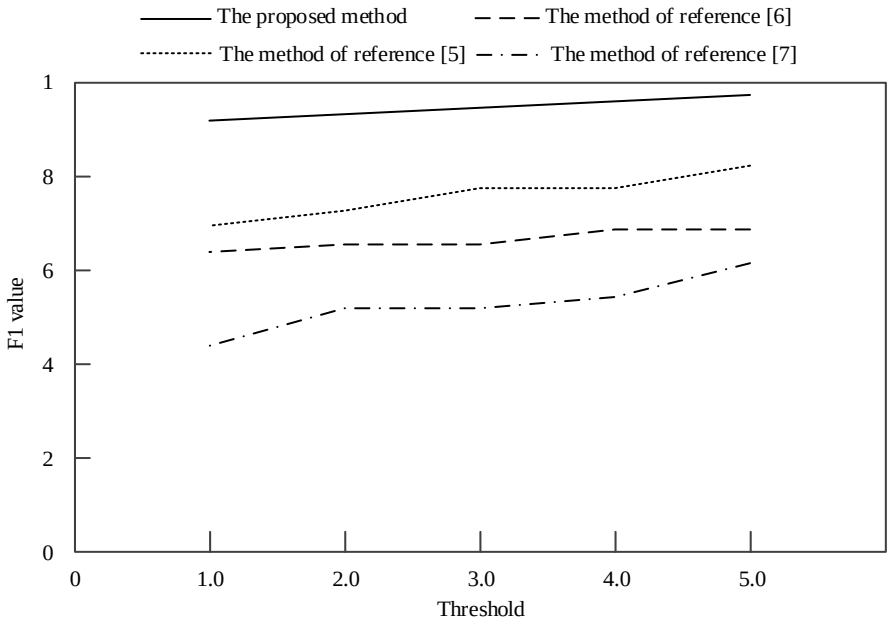

Figure 6. F1 Value Comparison Test

As shown in Figure 6, as the threshold continues to increase, the F1 value of the proposed method is always higher than 9, while the F1 values of the methods of reference (Van Nguyen *et al.*, 2020), (Wang *et al.*, 2022), and (Peng *et al.*, 2019) are around 8, 6, and 5, respectively, which are lower than the proposed method. According to the test results, the proposed method has the highest F1 value and can achieve high-precision mining of customer demand data.

6. CONCLUSIONS

Nowadays, with the continuous development of computer technology, the mining of customers' needs is becoming increasingly scarce, which leads to enterprises being unable to effectively mine customer demand information data. Therefore, how to reasonably and effectively mine customer demand information has become a very important research topic in the current era. Aiming at the problems of low mining feasibility and poor mining performance in current customer-demand mining, this paper proposes research on customer-demand mining algorithms based on online reviews and machine learning.

Using online comments as a data source for customer demand mining has broad coverage and naturalness, which can more comprehensively reflect customers' real needs and preferences, providing rich real-time and non-requested customer feedback for customer demand analysis. By eliminating duplicate and irrelevant data, the efficiency and quality of data mining can be improved. The processing of redundant information is an important part of data preprocessing, which is crucial for the accuracy and effectiveness of subsequent analysis and can lay a solid foundation for subsequent data mining. Extracting customer demand attribute features from processed data not only requires technical accuracy but also requires a deep understanding of customer needs to capture subtle differences and potential preferences. Using SOM (self-organizing map) neural network to cluster customer demand attribute features, SOM neural network has unique advantages in processing high-dimensional data and discovering the internal structure of data, which can help enterprises better understand the demand characteristics of different customer groups and provide strong support for precision marketing and product development. Thus, the design of a customer demand mining algorithm based on online comments and machine learning is completed. The experimental results show that the feasibility index of this method is as high as 90, customer satisfaction is always above 80, mining time is less than 3.4 seconds, and the F1 value is higher than 9. This method can effectively solve the problems in traditional methods and provide an important foundation for customer demand mining algorithms. However, with the development of future artificial intelligence, machine learning will become more intelligent. With the increasing complexity and diversity of customer needs, it is necessary to mine customer needs more accurately. Therefore, in the next research, the proposed method should be further improved to capture the innovative needs of customers, continuously improve the performance of data mining, and mine the needs of customers with greater accuracy.

REFERENCES

Dai, Y., Sun, X., Qi, Y. and Leng, M. (2021). A Real-Time, Personalized Consumption-Based Pricing Scheme for the Consumptions of Traditional and Renewable Energies. *Renewable Energy*, 180(C):452-466.

Durowoju, O., Chan, H.K. and Wang X. (2020). Investigation of the Effect of e-Platform Information Security Breaches: A Small and Medium Enterprise Supply Chain Perspective. *IEEE Transactions on Engineering Management*, 69(6):3694-3709.

Wang, T. and Zhou, M. (2021). Integrating Rough Set Theory with Customer Satisfaction to Construct A Novel Approach for Mining Product Design Rules. *Journal of Intelligent & Fuzzy Systems*, 41(1):331-353.

Guney, S., Peker, S. and Turhan, C. (2020). A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining. *IEEE Access*, 8: 84326-84335.

Van Nguyen, T., Zhou, L., Chong, A.Y.L., Li, B. and Pu, X. (2020). Predicting Customer Demand for Remanufactured Products: A Data-Mining Approach. *European Journal of Operational Research*, 281(3):543-558.

Wang, X., Lv, T. and Fan, L. (2022). New Energy Vehicle Consumer Demand Mining Research Based on Fusion Topic Model: A Case in China. *Sustainability*, 14(6):1-13.

Peng, N., Xie, Y., Li, Y., Wen, H., Dai, D. and Subinur. (2019). What's Your Ideal Online Short-Term Accommodation? Demand Mining for Chinese Tourists//2019 8th International Conference on Industrial Technology and Management (ICITM). *IEEE*, 369-373.

Liu, D., Huang, X. and Huang, K. (2020). Product Customer Demand Mining and Its Functional Attribute Configuration Driven by Big Data. *International Conference of Pioneering Computer Scientists, Engineers and Educators*, 145-165.

Lin, Q. and Son, J. (2020). A Data Mining Algorithm to Gaining Customer Loyalty to Ports Based on OD Data for Improving Port Competitiveness. *Journal of Navigation and Port Research*, 44(5):391-399.

Rozanec, J.M. (2021). Explainable Demand Forecasting: A Data Mining Goldmine. *Companion Proceedings of the Web Conference 2021*, 723-724.

Zare, M., Shakeri, H. and Mahmoudi, R. (2020). Ecommerce: An Efficient Digital Marketing Data Mining Framework to Predict Customer Performance. *Journal of the International Academy for Case Studies*, 26(5):1-8.

Arif, S.A. and Hossain, T.B. (2021). Opinion Mining of Customer Reviews Using Supervised Learning Algorithms//In 2021 5th International Conference on Electrical Information and Communication Technology (EICT). *IEEE*, Dec(17):1-6.

Moazzam, A., Mushtaq, H., Sarwar, A., Idrees, A., Tabassum, S. and Rehman, K.U. (2021). Customer Opinion Mining by Comments Classification using Machine Learning. *International Journal of Advanced Computer Science and Applications*, 12(5):385-393.

Zhang, J., Wang, C., and Chen, G. (2020). A Review Selection Method for Finding an Informative Subset from Online Reviews. Informs Journal on Computing, 33(1):280-299.

Liu, Y., Xiong, K., Lu, Y., Ni, Q., Fan, P. and Letaief, K.B. (2021). UAV-aided Wireless Power Transfer and Data Collection in Rician Fading. *IEEE Journal on Selected Areas in Communications*, 39(10):3097-3113.

Kauffmann, E., Pera, J., Gil, D., Ferrández, A., Sellers, R. and Mora, H. (2020). A Framework for Big Data Analytics in Commercial Social Networks: A Case Study on Sentiment Analysis and Fake Review Detection for Marketing Decision-Making. *Industrial Marketing Management*, 90:523-537.

Traub, J., Grulich, P.M., Cuéllar, A.R., Bre, S. Katsifodimos, A., Rabl, T. and Markl, V. (2021). Scotty: General and Efficient Open-source Window Aggregation for Stream Processing Systems. *ACM Transactions on Database Systems*, 46(1):1-46.

Che, Z., Borji, A., Zhai, G., Ling, S. Li, J., Tian, Y., Guo, G. and Le Callet, P. (2021). Adversarial Attack Against Deep Saliency Models Powered by Non-Redundant Priors. *IEEE Transactions on Image Processing*, 30:1973-1988.

Yang, G.H. and Feng, J.K. (2021). Network Intrusion Data Mining Simulation Based on Improved Apriori Algorithm (IAA) .*Computer Simulation*, 38(7):286-289+303.

Alsenan, S.A., Alturaiki, I.M. and Hafez, A.M. (2020). Auto-KPCA: A Two-Step Hybrid Feature Extraction Technique for Quantitative Structure-Activity Relationship Modeling. *IEEE Access*, 9:2466-2477.

Zhang, J., Chen, Y. and Zhai, Y. (2020). Zero-Shot Classification Based on Word Vector Enhancement and Distance Metric Learning. *IEEE Access*, 8:102292-102302.

Kim, T.G., Lee, Y.R., Kang, B.J. and Im, E.G. (2019). Binary Executable File Similarity Calculation Using Function Matching. *The Journal of Supercomputing*, 75(2):607-622.

Zhang, B., Cai, H., Chen, J., Hu, Y., Huang, J., Rong, W., Weng, W., Huang, Q., Wang, H. and Peng, H. (2020). Fast and Accurate Clustering of Multiple Modality Data via Feature Matching. *IEEE Transactions on Cybernetics*, 52(6):5040-5050.

Liu, X., Wang, T., Ji, T., Wang, H., Liu, H., Li, J. and Chao, D. (2022). Using Machine Learning to Screen Non-Graphite Carbon Materials Based on Na-Ion Storage Properties. *Journal of Materials Chemistry A*, 10(14):8031-8046.

He, B. and Yin, L. (2021). Prediction Modelling of Cold Chain Logistics Demand Based on Data Mining Algorithm. *Mathematical Problems in Engineering*, 2021(5):1-9.

Akbar, C., Li, Y. and Sung, W.L. (2021). Machine Learning Aided Device Simulation of Work Function Fluctuation for Multichannel Gate-All-Around Silicon Nanosheet MOSFETs. *IEEE Transactions on Electron Devices*, 68(11):5490-5497.

Kosasih, E.E. and Brintrup, A. (2022). A Machine Learning Approach for Predicting Hidden Links in Supply Chain with Graph Neural Networks. *International Journal of Production Research*, 60(17):5380-5393.

Zheng, J. and Ma, R. (2021). Analysis of Enterprise Human Resources Demand Forecast Model Based on SOM Neural Network. *Computational Intelligence and Neuroscience*, 2021(5):1-10.

Wang, X., Wan, T., Yang, Q., Zhang, M. and Sun, Y. (2021). Research on Innovation Non-Equilibrium of Chinese Urban Agglomeration Based on SOM Neural Network. *Sustainability*, 13(17):1-21.

Liu, Z., Wen, T., Sun, W. and Zhang, Q. (2020). Semi-Supervised Self-Training Feature Weighted Clustering Decision Tree and Random Forest. *IEEE Access*, 8:128337-128348.

Wang, Z., Chen, S., Guo, R., Li, B. and Feng, Y. (2020). Extreme Learning Machine with Feature Mapping of Kernel Function. *IET Image Processing*, 14(11):2495-2502.

Memis, S., Enginoglu, S. and Erkan, U. (2021). Numerical Data Classification via Distance-Based Similarity Measures of Fuzzy Parameterized Fuzzy Soft Matrices. *IEEE Access*, 9:88583-88601.