Industry Science and Engineering Vol. 1 No. 4, 2024

Clustering Analysis of Hotel Network Reviews Based on Text Mining Method

Yao Wang[1], Fuguo Liu[1,*], Guodong Li[2]
[1]School of Mathematics and Data Science, Changji College, Changji, Xinjiang, China,
[2]School of Mathematics and Computational Science, Guilin University of Electronic Science and Technology, Guilin, Guangxi, China
**Corresponding Author.*

Abstract: With the development of information technology, users use online platforms to post real-time online comments to express their preferences and opinions on goods or services. Online review information expresses users' behavioral habits and special preferences. In depth, analysis of hotel online reviews can improve the adaptability of hotel services to user needs. Effective mining of the vast user review data will provide value for the development of the tourism industry. Using text mining methods to process hotel review data, multiple clustering methods were compared and analyzed for positive and negative feature words from the perspective of user experience. It was found that the k-means++ algorithm had a better clustering effect on user network reviews and achieved better clustering and segmentation of user evaluation information. Unsupervised clustering analysis can be used to further classify online comment information into categories based on positive and negative reviews, providing intellectual support for improving the precision and personalized service quality of hotels.

Keywords: Network Comments; Text Mining; Cluster Analysis; Hotel; User Experience

1. Introduction

With the rapid development of tourism and the popularization of Internet applications, more and more users rely on online platforms to obtain travel information, book online hotels and express opinions and comments. According to the 53rd Statistical Report on China's Internet Development, as of December 2023, the number of users who book hotels, air tickets and other services online has reached 509 million, an increase of 20.4% year-on-year. The network behaviors of huge online users generate a large amount of user data, and the analysis and mining of these user data will provide new value for the development of online tourism [1]. In recent years, China has made great efforts to promote high-quality development, vigorously develop the digital economy, attach great importance to the digital economy, promote the growth of new technologies, new industries and new forms of business, promote the development of sharing economy with institutional innovation, and build a new engine with strong momentum [2]. Hotel service industry and information industry show coupling development, which promotes the development of sharing economy. Consumers' preferences for hotel products, access to information channels, evaluation, selection and final evaluation of hotel products are mostly completed through the network, and online reviews are an important way for consumer users to express their preferences and views immediately. User evaluation feedback has special commercial value, which can provide users with better hotel selection opportunities and recommendation guidance, but also attract new potential users to join the consumer group. Through feedback information, merchants can better understand the real needs of users, make reasonable market decisions in a timely manner, and enhance the value of goods by re-designing products to attract more online users to participate in the selection of services. Mining hotel online review data is of particular significance for tourism enterprises to improve service quality and expand tourism market space.

With the development of Internet technology, online comments have become an important way for netizens to express their opinions on

various industries, businesses, products and services. As one of the important contents of tourism, the comments of Internet users on hotels have gradually attracted the attention of researchers in recent years. From the perspective of users expressing comments, Maria P et al. [3] used machine learning algorithms to conduct sentiment analysis on hotel configuration, service and other aspects of hotel review data, and proposed a BiLSTM neural network to build a pre-training model for experiments. The classification effect is better based on traditional algorithms, which has certain reference significance for domestic hotel review research. However, there are some disadvantages such as single sample type and insufficient breadth of sample. Based on the hotel online review data provided by tourism websites, Qi Z et al. [4] analyzed the emotional tendency of various features of alternative hotels and considered the correlation of different features to recommend and rank alternative hotels, which helped tourists make decisions on hotel selection, but did not pay much attention to the needs of users. Tang Y L [5] et al. mined the online opinion review data in the hotel booking platform and used the fusion features of BiLSTM model to train the opinion data, so as to mine the opinion data in the comments more accurately.

Based on the text clustering method, Ji R [6] et al proposed a BERT language model based on pre-training and LDA topic clustering method to effectively mine online hotel reviews, classify the emotion of the text data through deep learning algorithm, dig out the problems that customers are deeply concerned about, and then put forward suggestions for managers. Aiming at the problem of poor text clustering performance, Shen X G [7] proposed a K-means text clustering algorithm with improved centroid initialization, which effectively solved the problem of local optimal results caused by the random selection of initial center points by traditional algorithms and made the algorithm more effective in text data mining. To sum up, online review analysis is gradually attracting the attention of researchers, who mostly use different feature words to carry out statistics and user portraits for hotel merchants but lack in-depth analysis of hotel reviews. In this paper, text mining method is used for text processing of online review information in hotel service industry, and different clustering algorithms are used for comparative analysis, so as to improve the efficiency of fine-grained cluster analysis of hotel online reviews, improve the ability of hotels to deeply analyze the characteristics of preferred groups of user experience, and promote merchants to expand their advantages and improve service quality.

2. Text Mining Related Methods

2.1 The Doc2vec Model

Word vector model is the result of classical language model neural network, the model on the model of training the statement every word can be generated at the same time the corresponding term vectors, and calculate the between words, and the correlation between paragraph text at the same time, more focus on the combination between words order. Doc2vec model [8] was proposed by TomasMikolov in 2014 based on the idea of Word2vec model. The basic idea of the Word2vec model is to predict the probability of the occurrence of the current word according to the context of the word. In this model, each word is mapped in a way into a vector, the mapped word vector is used as the input matrix of W, and the columns of the matrix are set by the word index in the phrase, and the occurrence of related words in the sentence is predicted by these word vectors. The goal of the word vector model is to predict the occurrence probability of unknown words according to the maximum value of the average logarithmic probability of known words, as shown in equation (1):

$$\frac{1}{N}\sum_{N=K}^{N-K}\lg p(w_n \mid w_{n-k},...,w_{n+k}) \quad (1)$$

$$p(w_n \mid w_{n-k},...,w_{n+k}) = \frac{e^{y_{wn}}}{\sum_i e^{y_i}} \quad (2)$$

Where, y_i is a non-normalized logarithmic function, and each word of the input layer is shown by formula (3).

$$y = a + bf(w_{n-k},...,w_{n+k};V) \quad (3)$$

a, b are SoftMax parameters, and f is calculated from the mean value of the word vector extracted from the word vector matrix V. Doc2vec model differs somewhat from the commonly used Word2vec method. Generally, Word2vec only trains word vectors, which is

somewhat imperfect for contextual semantic analysis. However, Doc2vec model adds a paragraph vector Paragraphid, which is equal in length to word vectors. It not only adds certain text semantic information, but also has specific text compatibility. Doc2vec has two common models: DBOW corresponding to skip-gram in Word2vec, DM corresponds to CBOW Word2vec, this paper adopts Doc2vec vectorization algorithm in the words of vector distribution of memory model (PV - DM), as shown in Figure 1.

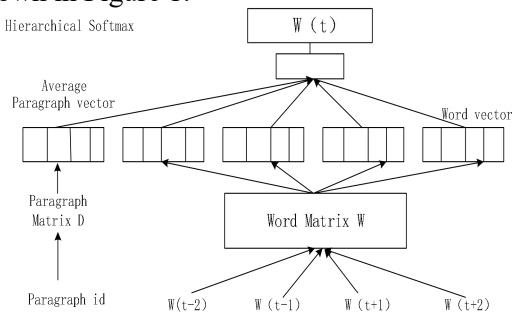

Figure 1. Doc2vec Model Description

The difference between the two models is that the PV-DM model predicts the target word based on the model in a given context, while the DBOW predicts the probability of the context based on the target word. In text mining, the PV-DM model first maps each text according to the specified dimension, and the column of matrix W is composed of the vector mapped into the context words of the target vocabulary. The target word is predicted by combining the paragraph vector and the word vector that can be used to store information. The sentence vector Paragraphvector and the word vector Wordvector have the same dimension, but they represent two different vector Spaces. The input of SoftMax layer is information that is accumulated or joined by Paragraphvector and Wordvector. To ensure the sharing of information in context Windows of a text, the random gradient descent model and backpropagation training method can be used to mine the text to obtain relevant features. In this paper, the Doc2vec model is used to calculate the text correlation among the studied web comments, which makes the text similarity calculation and processing of web comments have a good effect.

2.2 TF-IDF Algorithm

TF-IDF algorithm is a calculation method based on text statistics, which is often used to evaluate the importance of a specific word to a document in a document set. The characteristic described by the TF-IDF algorithm is that the more important a word is to the document, the more likely it is to be the key word of the document.

Tf-idf algorithm consists of two parts: TF algorithm and IDF algorithm. The TF algorithm counts how often a word appears in a document. The more times a word appears in a document, the better it expresses the document. The IDF algorithm counts how many documents a word appears in a document set. The basic idea is that the fewer documents a word appears in a document collection, the better its keyword will be at distinguishing documents. The calculation formula is as follows:

$$tf_{ij} = \frac{n_{ij}}{\sum_k nk_j} \tag{4}$$

The numerator is the number of occurrences of the word in the file, and the denominator is the sum of all occurrences of the word in the file.

$$idf_i = \log(\frac{|D|}{1+|D_i|}) \tag{5}$$

Where: $|D_i|$ is the total number of documents in the document set, $|D_i|$ is the number of documents where the word i appears in the document set. Adding 1 to the denominator uses Laplacian smoothing to avoid the situation that may cause the denominator to be zero and enhance the robustness of the algorithm. TF-IDF algorithm is a comprehensive use of the two methods, and the calculation formula is as follows.

$$tf - idf(i,j) = tf_{ij} * idf_i$$
$$= \frac{n_{ij}}{\sum_k n_{kj}} * \log(\frac{|D|}{1+|D_i|}) \tag{6}$$

2.3 T-SNE Algorithm

T distribution stochastic neighborhood embedded called T - SNE algorithm is a kind of typical dimension reduction algorithm [9], it was proposed in 2008 by Rene Vander Maaten and Geoffrey Hinton and has attracted much attention from researchers in various fields in recent years. The basic idea of T-SNE algorithm is to observe the joint distribution of

multiple feature data points, use the corresponding similarity recognition mode to divide the data into the corresponding clusters, find the corresponding embedding mapping relationship, and map the points in the high-dimensional space to the low-dimensional space, while keeping the probability of distribution between each other unchanged. In low dimensional space, T distribution is used to convert distance into probability distribution, and specific sample distribution is obtained by distance KL divergence, which can achieve good reduction and data visualization analysis. Set the high-dimensional data points as $X=(X_1,X_2,X_3,...,X_n)$, the low-dimensional data points are $Y=(Y_1,Y_2,Y_3,...,Y_n)$, The KL divergence is calculated as follows:

$$C=KL(P\left\|Q\right)=\sum_i\sum_j p_{ij}Lg\frac{p_{ij}}{q_{ij}} \quad (7)$$

In the above formula, p_{ij} is the joint probability function of the sample distribution in a high-dimensional space, and q_{ij} is the joint probability of the sample distribution in a low-dimensional space, then the conditional probability of the high-dimensional data is as follows:

$$P_{j|i}=\frac{\exp(-\left\|x_i-x_j\right\|^2/2\sigma_i^2}{\sum_{i\neq j}\exp(-\left\|x_i-x_j\right\|^2/2\sigma_i^2)} \quad (8)$$

The value in the above equation is the Gaussian variance centered on the value, and the optimal value is obtained through a binary search of the pre-set complex factors. A symmetric joint probability can be defined by the following formula:

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n} \quad (9)$$

The joint probability with symmetry in a low-dimensional data space can be described by:

$$q_{ij}=\frac{(1+\left\|y_i-y_j\right\|^2)^{-1}}{\sum_{i\neq j}(1+\left\|y_i-y_j\right\|^2)^{-1}} \quad (10)$$

In addition, the gradient of optimization can be described by the following formula:

$$\frac{\partial c}{\partial Y_i}=4\sum_i(p_{ij}-q_{ij})(y_i-y_j)(1+\left\|y_i-y_j\right\|^2)^{-1} \quad (11)$$

T-SNE algorithm defines the specific boundary features between the local and global structure of data. This algorithm is a popular dimensionality reduction method at present, which provides theoretical support for data visualization.

3. Algorithm Simulation

In the era of digital economy, text mining technology is applied to the analysis and research of hotel online reviews, and reasonable cluster analysis is carried out on user reviews to mine potential information for merchants to adjust marketing strategies, which is conducive to merchants to provide accurate services or recommendation information. Cluster analysis methods usually group and classify seemingly disordered objects, gather the same type of data together, the similarity of objects within the group is high, and the similarity of objects between groups is low, separate the data with large differences as far as possible, and analyze the data characteristics.

4. Clustering Algorithm

4.1 k-means Algorithm

In the traditional k-means algorithm, the sample set D is divided into several disjoint subsets through iteration technology, and the cluster class is divided according to the distance between the sample and the cluster center point, and then iterated until convergence [10]. The final goal is to minimize the square error, and the objective function is:

$$E=\sum_{i=1}^{K}\sum_{x\in c_i}\left\|x-\mu_i\right\|^2 \quad (12)$$

The traditional k-means clustering algorithm needs to set cluster k in advance, and its value is difficult to estimate. In many cases, a given data set cannot be determined to cluster into several classes in advance, and the time complexity of the algorithm is relatively influenced by class clusters.

4.2 Canopy k-means Algorithm

Canopy algorithm pre-classifies data on the basis of K-means algorithm and can be approximated by the number of large circles generated by canopy when the number of cluster centers cannot be determined artificially. Canopy processes data sets through two subjectively determined thresholds t1 and t2, which can sort chaotic data into several data

piles with certain rules. To solve the problem that the number of classifications is difficult to determine, the Canopy K-means algorithm can approximate the number of cluster centers by the number of large circles generated by canopy. However, in the actual application of the algorithm, the selection of initial Canopy center point and Canopy area size and other initial values has a great influence on the clustering quality.

4.3 k-means++ Algorithm

k-means++ algorithm [11] is an improved unsupervised clustering algorithm based on the traditional K-means algorithm. The basic idea is to take the distance from the data point to the cluster center as the objective function of optimization and obtain the optimization rule of iterative operation by using the method of function extremum. The specific steps are: Accept the cluster number k specified by the user, randomly select the first cluster center. When selecting the n ($n \in [2,k]$)cluster center, the sample farther away from the first N-1 cluster center has a higher probability of being selected, so as to select k initial cluster centers, and assign the sample to the category closest to the cluster center through iteration. At the same time, the average value of each type of sample is calculated as the new clustering center, and the objective function of the algorithm is finally minimized. Its objective function is shown in equation (13).

$$E = \sum_{i=1}^{k} \sum_{j=1}^{N} u_{ij} d_{ij} \tag{13}$$

Where: k is the number of clusters, N is the number of samples in the sample set, u_{ij} is the membership of sample j to class $u_{ij=1}$ is that sample j belongs to class i, $u_{ij} = 0$ is the opposite, d_{ij} is the distance between the cluster centers of sample j and class i samples. The goal of the algorithm is to distance the clustering centers from each other as far as possible. Although this improvement is intuitive and simple, it is very effective, and better improves the randomness of the initial center point selection of K-means algorithm.

5. Empirical Analysis

The experimental environment was built on the platform of Core i3720M (1.80GHz) processor, 4.00GB memory, 64-bit Windows10 operating system, and python3.6 was used as the development language for algorithms.

5.1 Hotel Online Review Data Source

This paper extracts the review data of hotel user experience in Urumqi from Meituan e-commerce platform by Scrapy in python software. In order to reflect the computational efficiency and visualization of the data, 2000 positive reviews above four stars and 2000 negative reviews below two stars are randomly extracted as the data set used in this paper.

The hotel user experience review information crawled from the online platform is basically text information, which belongs to non-numerical data. In order to compare the similarity between subsequent documents, it is necessary to convert non-numerical data such as text information into numerical data and classify according to the size of similarity. In recent years, many researchers have written a library of relevant algorithms for the Python computer design language, which has formed a language ecological library for big data analysis, which is convenient for users to call python packages for program design.

In this paper, the classic Jieba Chinese word segmentation tool in Python software is used to process the hotel text data, and the processed corpus is stored in the TXT document. After word segmentation, this paper filters the data set with invalid terms as reference and deletes the words with only one word in the results of word segmentation to reduce the data dimension and noise to a certain extent. Finally, the remaining words are separated by Spaces in the order in which they appear for data storage.

Text representation is the basic work of natural language processing, and the quality of text representation directly affects the performance of the whole natural language processing system. Therefore, vectorization algorithm of doc2vec model in Gensim library was adopted in this paper to conduct paragraph vector training on corpus data. Parameters of vector model are shown in Table 1. The trained model is used to perform vectorization operations on each text, and the similarity of comments is calculated for 2000 positive and negative comments respectively, and two multidimensional (2000*2000) similarity matrices are obtained. The popular T-SNE

algorithm uses the T-distribution random neighborhood embedding mechanism to reduce the dimensionality of the multidimensional matrix and maps the multidimensional data to the two-dimensional data space, which is convenient for visual analysis of the data results.

Table 1. Doc2vec Parameter Settings

argument	value
Size (dimension of sentence vector)	100
Window (window length)	10
Min-count (minimum number of occurrences)	2
Workers (Thread count)	2

5.2 Cluster Analysis of Hotel Online Reviews

In this paper, three clustering algorithms were used to measure the data set, namely, traditional k-means algorithm, canopyk-means algorithm and K-Means ++ algorithm. Before the cluster analysis, feature words of the text are first selected. The TF-IDF algorithm is used to select words with greater weight from the two categories of positive and negative comments as feature words for comments, and the words are arranged according to the order of weight from large to small, as shown in Table 2 and Table 3.

Table 2. The Weight of "Favorable" Feature Words

WORD	TFIDF	WORD	TFIDF
right	96.2575	enthusiasm	16.0971
expediency	51.4919	Service attitude	15.5240
facility	38.3670	Queen bed	14.7959
neat	33.4784	hardware	14.4599
guesthouse	32.3466	Railway station	12.7700
traffic	31.8084	Sound insulation	12.5516
position	30.3378	comfort	12.5352
Geographical position	23.2728	luxury	12.4342
satisfaction	18.3718	Neat and tidy	12.2415
comfy	16.3960	quiet	12.1608

The topic distribution trained by TF-IDF model was applied to the review data, and the relevant feature words were obtained by setting topic=20. As shown in Table 2 and Table 3, in the category of hotel praise, it can be seen that the top keywords are good, service and convenience. Customers' evaluation of the hotel is more favorable to the service provided by the hotel and the geographical location of the hotel. Secondly, users' evaluation pays attention to the perception, that is, the first impression. Finally, customers will carefully evaluate the details of the sleep environment, the core function of the accommodation, such as decoration, large bed, soundproofing, comfort, reflecting the diversified service needs of customers. As can be seen from the bad rating score table in Table 3, the details of user evaluation are diversified and more focused on user experience. After staying in, users focus on decoration, air conditioning, toilet, taste, elevator, Internet access, towels, Windows, warm water and quilts. Compared with Table 3, Table 2 has more favorable macro classification with more dimensions and wide coverage, while Table 3 has more detailed feature extraction.

Table 3. Weight Value of "Bad Evaluation" Feature Words

WORD	TFIDF	WORD	TFIDF
obsolescence	22.7948	carpet	11.0352
air conditioner	20.7895	feedback	10.5408
Rest room	19.7499	window	10.4903
Check out	18.3463	Hot water	10.0626
Room rate	13.5196	Musty smell	9.0143
disappointment	13.4393	Quilt	7.6054
taste	13.1416	toilet	7.0519
Surf the Internet	11.6839	heating	6.4428
Surf the Internet	11.5607	Parking lot	6.1176
Cost performance	11.2940	Peculiar smell	6.0436

After processing the text review data, the traditional k-means algorithm, canopyk-means algorithm and K-Means ++ algorithm are used to cluster the two-dimensional data sets of positive and negative reviews respectively. The comparison results of experimental performance are shown in Table 4.

According to the comprehensive analysis of poor rating data and good rating data, the clustering effect of Canopyk-means algorithm has a small improvement compared with the traditional k-means algorithm, while the k-means++ algorithm has a small improvement compared with the traditional K-means algorithm and Canopyk-means algorithm. The contour coefficient and BWP index have a

large improvement, because the K-Means ++ algorithm can reliably obtain the best cluster number, the result has a small deviation from the actual cluster number. K-means++ clustering algorithm is more effective than other algorithms in analyzing text data.

Table 4. Comparison of Experimental Results of Different Clustering Methods

Clustering model	(Negative comment) Contour system	(Negative comment) BWP	(Good reputation) Contour system	(Good reputation) BWP
k-means algorithm	0.0499	0.2651	0.0559	0.3195
canopyk-means algorithm	0.0643	0.3392	0.0681	0.4229
K-Means ++algorithm	0.0920	0.6684	0.1025	0.6521

In Figure 2 (a), the traditional k-means clustering effect is insensitive to samples and lacks feature recognition of the overall samples. The clustering effect of K-Means ++ in (c) is slightly improved compared with the clustering effect of canopyk-means in (b). Compared with other clustering algorithms, the intra-cluster tightness is higher, indicating that the clustering effect is better, and the clustering characteristics are significant. In Figure 3 (a), the traditional k-means clustering effect is poor and lacks substantial response to the overall sample. In Figure. (c), the clustering effect of K-Means ++ is obvious and the clustering effect is better than that of canopyk-means clustering effect. Based on Figure. 2 and Figure. 3, it can be observed that the K-Means ++ algorithm has a good overall evaluation on the clustering effect of user praise and user negative evaluation, while the canopyk-means algorithm has no significant clustering effect but is suitable for the discovery of detailed features.

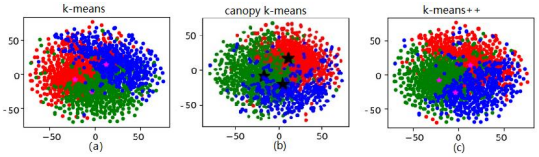

Figure 2. Clustering Results of "Favorable" Comments by Different Clustering Methods

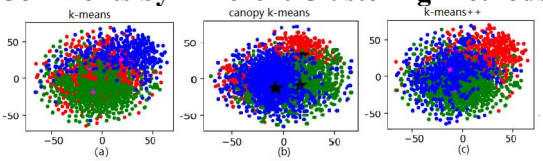

Figure 3. Clustering Results of Different Clustering Methods on "Bad Reviews"

The K-Means ++ clustering algorithm was used to cluster the reviews, and three kinds of cluster results were obtained. The first category has 693 reviews, focusing on hotel hygiene and general reviews on the environment. The second category has 781 comments, focusing on the evaluation information of hotel infrastructure construction; the third category, with a total of 526 reviews, focuses on the hotel's service evaluation information.

In the same way, the cluster analysis of bad reviews can be divided into three types of cluster results. The first type has 630 reviews, focusing on the evaluation information of hotel infrastructure construction. The second category has 521 reviews, focusing on information about the external environment and sanitary conditions of the hotel; the third category has a total of 849 reviews and focuses on the evaluation information of hotel services. According to the cluster subdivision theme, from the word cloud map, users focus on the health environment and infrastructure construction of the hotel. From the user evaluation, "praise" users pay more attention to the hotel's service attitude, focusing on the spiritual level of satisfaction; The difference is that from the feedback of "bad reviews", users focus on the sanitary environment of hotel facilities, from the basic conditions and demand details to put forward feedback information, therefore, the hotel industry needs to improve the hotel environment from the details to meet the needs of users.

6. Result Discussion

6.1 Detailed User Evaluation Scenarios

Among the positive and negative comments, it is found that user comments cover online booking, check-in procedures, hotel going out for leisure, use of indoor facilities, check-out procedures, etc. In the analysis of positive comments, "satisfaction" and "room" appear more frequently, and "good" and "convenient" are habitual words of most users. However, users use "good" and "convenient" to evaluate, relying on "service", "environment", "facilities" and "price", while "room" can be regarded as the focus of most user evaluations.

In the case of user preference, more attention is paid to the external situation of hotel facilities, such as traffic, geographical location, etc., while in the case of user aversion, more attention is paid to the indoor decoration and configuration of hotel facilities, such as towels, carpets, etc.

6.2 User Experience Involves Multiple Subjects

The factors affecting user experience involve multiple subjects, including hotel facilities, surrounding environment and services. Hotel facilities include in-store facilities and out-of-store facilities, in-store facilities include beds, decoration, Internet access equipment, stairs, etc. These experiable elements are expressed through the user's sense of smell, touch, psychological feelings and vision, such words include "taste", "old", "musty" and "disappointment", while out-of-store facilities include parking Spaces, etc. The surrounding environment involves convenient transportation and location to meet the diversified needs of users, while the hotel attendants and users can communicate directly, so it brings intuitive feelings to users, and the word "attitude" is an important evaluation index of users on the attendants.

6.3 Hotel Business Services "Short Board" Highlights

As shown in Table 2 and Table 3, "soundproof" and "large bed" belong to the "short board" content. "Soundproof" and "big bed" ranked lower in Table 2 and higher in Table 3, it can be seen that there are shortcomings in the two contents of rest quality and sleep quality, obviously the hotel industry's attention to the sleep quality of users is not enough, hotel businesses can adjust the strategic layout force, need to focus on making up for these shortcomings.

6.4 Create an Atmosphere for Continuous Improvement of the User Experience

The extraction of hotel network praise and bad review words can help hotel merchants understand the consumption habits of mainstream user groups and form a positive interactive feedback mechanism for users and merchants. It is convenient for businesses to gain insight into users' consumption needs from the micro level, capture the "pain points" of user experience, enhance users' trust in business products, and cluster analysis of network comments will help hotel businesses tap into the subdivision of user experience preferences and create an atmosphere for continuous improvement of user experience.

7. Conclusion

In this paper, Doc2vec is used to shorten the dimension of feature vector to extract text context information, and TF-IDF is used to label each class to realize comment text mining. This paper captured hotel review information from the Meituan platform as a dataset and divided the positive subsets and the negative subsets according to star rating for feature analysis. Compared with the negative subsets, the positive subsets are more macro classification, with more dimensions and wide coverage, while the negative subsets are rich in detailed feature extraction, and the extracted feature words are more focused on user experience. In order to further cluster analysis of review data, this paper adopts the traditional k-means algorithm, Canopyk-means algorithm and K-Means ++ algorithm to conduct a comparative study. Through clustering comparison of the positive and negative subsets, it is found that k-means++ is superior to Canopyk-means algorithm and traditional K-means algorithm in aggregation degree. Through the cluster analysis of network comments, the user experience of the hotel is continuously refined. In the case of positive preference, relying on "service", "environment", "facilities" and "transportation" left users with a sense of "satisfaction"; In the case of diversified user experience, the location, environment and traffic of the hotel directly affect user satisfaction; In the case of negative aversion, because the service quality of hotel merchants is more transparent on the platform, merchants pay less attention to sleep quality. The cluster analysis of hotel online reviews based on text mining can help merchants accurately play their advantages and make up for their shortcomings, improve the overall service quality of the hotel industry, and expand the space for tourism development.

References

[1] Yang L, Maomao C, Qiong S. Sarcasm detection in hotel reviews: a multimodal deep learning approach. Journal of

Hospitality and Tourism Technology, 2024, 15(4): 519-533.
[2] Krishnan J, Bhattacharjee B, Pratap M, et al. Survival strategies for family-run homestays: Analyzing user reviews through text mining. Data Science and Management, 2024, 7(3): 228-237.
[3] Maria P, John G, Bay K O. Consumer-brand heuristics in luxury hotel reviews. Journal of Product & Brand Management, 2024, 33(4): 436-448.
[4] Qi Z, Yuling L, Hang D, et al. Public concerns and attitudes towards autism on Chinese social media based on K-means algorithm. Scientific Reports, 2023, 13(1): 15173-15173.
[5] Tang Y L, Wang H Z, Wang D S, et al. A novel rough semi-supervised k-means algorithm for text clustering. International Journal of Bio-Inspired Computation, 2023, 21(2): 57-68.
[6] Ji R, Yang J, Wu Y, et al. Correction to: Construction and analysis of students' physical health portrait based on principal component analysis improved Canopy-K-means algorithm. The Journal of Supercomputing, 2024, 80(14): 21561-21562.
[7] Shen X G, Jiang Y Z. Optimisation of K-means algorithm based on sample density canopy. International Journal of Ad Hoc and Ubiquitous Computing, 2021, 38(1-2-3): 62-69.
[8] Liang Z, Yiqu Z, Yumeng W, et al. Unveiling Patterns and Colors in Architectural Paintings: An Analysis by K-Means++ Clustering and Color Ratio Analysis. Tehnički vjesnik, 2023, 30(6): 1870-1879.
[9] Singh A, Koju R. Healthcare Vulnerability Mapping Using K-means ++ Algorithm and Entropy Method: A Case Study of Ratnanagar Municipality. International Journal of Intelligent Systems and Applications (IJISA), 2023, 15(2): 43-54.
[10] Li S, Chen J. Virtual human on social media: Text mining and sentiment analysis. Technology in Society, 2024, 78102666-102666.
[11] Zolfaghari B, Bibak K, Koshiba T, et al. Statistical Trend Analysis of Physically Unclonable Functions: An Approach via Text Mining. CRC Press: 2021-01-12.