Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9905114, 13 pages
https://doi.org/10.1155/2022/9905114

Research Article

Mining and Application of Tourism Online Review Text Based on Natural Language Processing and Text Classification Technology

Hongsheng Xu[1,2] **and Yanqing Lv**[1,2]

[1]*College of Electronic Commerce, Luoyang Normal University, Luoyang, 471934 Henan, China*
[2]*Henan Key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang Normal University, Luoyang, 471934 Henan, China*

Correspondence should be addressed to Hongsheng Xu; xhsls@lynu.edu.cn

Received 1 April 2022; Revised 2 May 2022; Accepted 8 May 2022; Published 23 May 2022

Academic Editor: Lei Zhang

This paper firstly describes the research status of online review text mining and finds out the problems existing in the mining and application of tourism texts. Aiming at these problems, this paper proposes a text mining method for tourism online reviews based on natural language processing and text classification technology. The first step is to analyze the validity of the online review text; the purpose is to remove the invalid text and improve the mining efficiency of the online review text. The second step is to conduct a comprehensive evaluation of scenic spots and hotels based on text classification technology and sentiment analysis. The comprehensive evaluation indicators are established for the five core service contents. High-quality scenic spots and hotels are selected according to the ranking of comprehensive evaluation. The third step is to propose a mining method of tourism hot words based on natural language processing for the selected high-quality tourist locations. The obtained hot words can intuitively show the impression of tourists on the scenic spot. The fourth step is to use mutual information combined with the left and right entropy to discover new words and to mine service characteristics of high-quality scenic spots and hotel from the new words. Finally, the proposed new methods are tested on the crawled tourism online review texts. The experimental results show that the novel comprehensive evaluation method proposed in this paper can truly and objectively select high-quality scenic spots and hotels and provide an important basis for the decision-making of tourism management. On this basis, hot words and new words can be effectively excavated from relevant online review texts, and travel impressions can be fed back from various aspects and angles.

1. Introduction

With the booming development of tourism, people pay more and more attention to practical experience when traveling. Relevant government departments and tourism enterprises are also focusing on improving the service quality of the tourism industry. Major travel websites, such as Trip, TravelGo, and eLong, all provide a wealth of tourist comment functions. Tourists can comment on scenic spots or hotels in tourist destinations from various aspects. These online review texts generally reflect the visitor's experience. These review texts can not only provide reference for other tourists but also provide suggestions for improving and enhancing tourism services for the operation department of the scenic spot, the local cultural tourism management department, and tourism enterprises [1].

Online review texts are often created by thousands or even tens of thousands of tourists and are a type of user generated content (UGC). In recent years, with the help of modern computers and artificial intelligence, especially natural language processing technology, these text data can be automatically processed and analyzed to obtain indicators reflecting the impression of tourists, thus providing a decision-making reference for the development of tourism [2]. At present, many scholars have carried out research on improving tourism services by analyzing online review texts.

This article uses the blog post about Zhujiajiao in Sina blog to divide the tourism perception image of Zhujiajiao from the direction of text mining [3]; Cai et al. [4] studied the audience perception of urban tourism image in Guiyang City based on ROST text mining software. This article used text mining method to conduct word frequency, sentiment, and semantic network analysis of online travel notes about Gansu tourist spots based on four typical travel websites such as Trip, Mafengwo, Lvmama, and Tuniu [5]. Li et al. [6] used text mining to conduct image perception research on the comments of tourists on typical urban tourism communities in Beijing, such as Baidu Travel, https://Trip.com/. This article conducted a comparative study of tourists' opinions and suggestions based on text mining in Guangxi Qinbeifang. This paper also formulated tourism policies for the government and related tourism management departments and provided an important direction for the development of tourism in Guangxi Qinbeifang and has important practical significance [7].

After analyzing the current research status, it is found that the above research is not systematic in the application of online review texts, lacks a comprehensive evaluation mechanism for scenic spots and hotels, and cannot reflect tourists' travel impressions from multiple angles and levels, so it cannot effectively feedback the services and characteristics of hotels and scenic spots. This paper proposes the analysis and processing of online review text based on natural language processing and text classification technology. The first is the validity analysis of online review text; the purpose is to remove invalid text. The second step is to use the text classification technology combined with the sentiment analysis method to carry out the comprehensive evaluation of the scenic spots and hotels, to comprehensively score the five services that the tourists are concerned about, and to select the high-quality tourist spots and hotels. The third step is to use the named entity recognition method in natural language processing to mine hot words for the top-ranked scenic spots and hotels. The obtained hot words can effectively feedback tourists' intuitive impression of scenic spots and hotels. The fourth step is to analyze the characteristic services of scenic spots and hotels based on new word discovery. The service characteristics of high-quality scenic spots and hotels are excavated in the new words. Therefore, these methods can reflect tourists' travel impressions from multiple perspectives.

2. Related Technical Analysis

2.1. Text Classification Techniques

2.1.1. Bayesian Methods and Naive Bayes. Bayesian methods and theories were first proposed by British mathematician Thomas Bayes. In recent years, with the development of artificial intelligence, especially the rise of machine learning, data mining, and other technologies, Bayesian theory has a broader development and application space.

The Bayesian classifier is a general term for a class of classification algorithms, which are all based on Bayes' theorem. Their applications in text mining are mainly focused on plain Bayesian classifiers and Bayesian network classifiers. The Naive Bayes is the simplest and most common type of Bayesian classifier. The algorithm assumes that the probability of all words appearing in the text is considered to be relatively independent [8].

Assuming that the set X is the set of text categories, determining whether a text y belongs to a category x_i can be done by calculating the probability of $P(x_i \mid y)$, i.e., given a text y, calculate what is the probability that it belongs to the text category x_i. The discriminant rule of plain Bayes is to categorize y into the category that makes $P(x_i \mid y)$ reach the maximum probability, i.e., to solve $\operatorname{argmax} P(x_i \mid y)$.

The Naive Bayes is the simplest and the most common of the Bayesian classifiers. The algorithm assumes that the likelihood of each attribute taking values is considered to be independent and uncorrelated with the values of other attributes [9]. The core idea is that for a given item to be classified, the probability of occurrence of each category under the conditions of this item is solved, and whichever is the largest is considered to be the category to which this item to be classified belongs.

Let X be an unlabeled data sample, and let H be some assumption that the data sample X belongs to a particular class C. In the classification problem, we want to obtain $P(H \mid X)$, i.e., given the predicted data sample d, the probability that the assumption H holds, and the classification is completed by comparing the maximum probability [10]. Its Bayes' theorem is formulated as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}, \tag{1}$$

where $P(H \mid X)$ is the posterior probability, which denotes the probability of occurrence of H given that condition X is found. $P(H)$ is the prior probability, which denotes the prior probability of hypothesis H. $P(X)$ is the prior probability of condition X. $P(H)$ is independent of X.

The process of Naive Bayes is as follows.

Each data sample is represented by an n-dimensional feature vector $d = \{x_1, x_2, x_3, \cdots, x_n\}$, which describe the n attributes $d_1, d_2, \cdots, d_n$ samples with n metrics. Assume m classes $c_1, c_2, \cdots, c_n$. Given an unknown data sample d, the classification method will predict that d belongs to the class with the highest posterior probability. That is, the unknown data sample is assigned to class c_i.

The class c_i whose $P(c_i \mid d)$ is the largest is called the maximum a posterior hypothesis and according to Bayes' theorem.

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)}. \tag{2}$$

Since $P(d)$ is constant for all classes, it is only necessary that $P(d \mid c_i)P(c_i)$ is maximal. After calculating $P(d \mid c_i)P(c_i)$ for each class, the sample d is assigned to the class c_i whose $P(d \mid c_i)P(c_i)$ is the largest.

For the calculation of the probability estimate, to improve its accuracy, the Laplace smoothing estimate can be used with the following equation.

$$P\left(c_j\right) = \frac{\sum_{i=1}^{|D|} P\left(c_j|d_i\right)}{|D|} (j = 1, 2, \cdots, |C|), \tag{3}$$

$$P\left(d_i|c_j\right) = \frac{1 + \sum_{t=1}^{|D|} B_{it} P\left(c_j|d_i\right)}{2 + \sum_{t=1}^{|D|} P\left(c_j|d_i\right)} (j = 1, 2, \cdots, |C|\,; t = 1, 2, \cdots, n), \tag{4}$$

where D is the training text set and $P(c_j \mid d_i) \in \{0, 1\}$, indicating whether the training text d_i belongs to the text of class c_j, which "1" means belonging, and "0" means not belonging.

2.1.2. Classifier of Linear Support Vector Machine. Support vector machine (SVM) algorithm is considered to be one of the more effective methods in text classification, which is a machine learning method based on statistical learning theory [11]. This technique solves the previous problem of requiring an infinite number of samples by simply abstracting a certain amount of text into vectorized training text data through computation, which improves the accuracy of the classification and can be widely used in statistical classification and regression analysis. It maps the vectors into a higher dimensional space in which a maximum interval hyperplane is established [12]. Two hyperplanes parallel to each other are built on either side of the hyperplane separating the data, and the separation hyperplane maximizes the distance between the two parallel hyperplanes. It is assumed that the greater the distance or gap between the parallel hyperplanes, the smaller the total error of the classifier.

Definition 1. For a given data set that is linearly divisible, this equivalent method of using interval maximization or solving the corresponding convex quadratic programming problem. The separation hyperplane is learned as $\omega^* x + b^* = 0$, and the corresponding classification decision function is $f(x) =$ sign $(\omega^* x + b^*)$. It is called linearly separable SVM.

Definition 2 (geometric interval). For a given training sample data set M and hyperplane (ω, b), we call the function interval of the hyperplane about the sample point (x_i, y_i) as $\gamma_i = y_i(\omega/\|\omega\| \cdot x_i + b/\|\omega\|)$. The minimum value of the geometric interval between the hyperplane (ω, b) and all sample points in the data set M is $\gamma = \min_{i=1,2,\cdots n} \gamma_i$.

SVM has few parameters, and the most important one is the kernel function. When there are relatively many text features, the linear kernel function is enough. The SVM using linear kernel function is called linear SVM; it is also the kernel function used in this paper.

SVM (support vector machine) is a statistical theory classification algorithm. The algorithm implements sample features to find a balance between model complexity and learning ability. It is more effective in solving small samples, nonlinearity, and high dimensionality.

The optimal classification plane $\omega \cdot \phi(x) + b = 0$ exists for a sample S. The original parameter space is transformed to a high-dimensional space using a nonlinear function, and then, the above hyperplane is established; ω denotes the plane normal vector, and b denotes the intercept set to separate the plane function:

$$\omega \cdot \phi(x) + b = 0. \tag{5}$$

The discriminant function $h(x) = \omega \cdot \phi(x) + b$; $h(x)$ will be normalized, so that the samples meet $h(x) \gg 1$; the operation transformation can be obtained after the simplification of the formula $D = 2/\|\omega\|$, so that the classification interval is the maximum is equivalent to make $\|\omega\|$ minimum.

To satisfy the above conditions, the following conditions are needed to make the classification models work correctly for all samples [13].

$$y_i(\omega \cdot \varphi(x_i) + b) - 1 \gg 0 (i = 1, 2, \cdots, n). \tag{6}$$

For each inequality constraint introduce a Lagrange multiplier (Lagrange multiplier) $\alpha_i \geq 0, N\alpha_i \geq 0, i = 1, 2, \cdots, N$; construct the Lagrange function:

$$L(\omega, b, a) = \frac{\|\omega\|^2}{2} - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\omega^T \varphi(x_i) + b\right) - 1\right). \tag{7}$$

By eliminating ω and b, the original constrained optimization problem can be equated to the minimax dual problem.

When linearly inseparable, it needs to be transformed to a higher dimensional space to make it linearly separable. In this case, the relaxation variable method is needed to solve this kind of problem, and a relaxation variable $\delta \geq 0$ is introduced for each sample so that the constraint becomes:

$$y_i\left(\omega^T \varphi(x_i) + b\right) \geq 1 - \delta_i. \tag{8}$$

The objective function then becomes:

$$\frac{\min \|\omega\|^2}{2} + C \sum_{i=1}^{i} \delta_i. \tag{9}$$

The larger C is the penalty factor, the smaller the number of error points, but it should not be too large to avoid excessive clutching. To avoid dimensional catastrophe, it is necessary to introduce the kernel function, which is a mapping from low-dimensional space to high-dimensional space.

2.2. Mutual Information and Left-Right Entropy Theory

2.2.1. Information Entropy. Information entropy was originally proposed by Shannon and can be used to describe the uncertainty of an event. Usually, a text with a lower probability of occurrence of an event contains more information and thus has a higher information entropy [14]. For example, "Bill Gates goes bankrupt" has a much lower probability of occurring than "Bill Gates becomes the richest man," so the former has a higher information entropy [15]. The formula of information entropy can be expressed as follows:

$$H = -\sum_{i=1}^{n} P_i \cdot \log_2(P_i). \tag{10}$$

2.2.2. Information Gain. Information gain (IG) (Mitchell1 1997) represents the average information of a document class when a text contains a certain feature, defined as the difference in information entropy before and after a feature appears in the text [16]. Assume that C is the set of text classes, c is the text class variable, d is the text, and t is the feature. For feature t, its information gain is denoted as $\mathrm{IG}(t)$. The frequency of documents with and without t occurring in c is examined. It is used to measure the information gain of word t for category c. The calculation formula is shown as follows:

$$\mathrm{IG}(t) = P(t)\sum_{i=1}^{|C|} P(c_i|t)\log\frac{P(c_i|t)}{P(c_i)} + P(\bar{t})\sum_{i=1}^{|C|} P\left(c_i|\bar{t}\right)\log\frac{P\left(c_i|\bar{t}\right)}{P(c_i)}. \tag{11}$$

2.2.3. Mutual Information. Mutual information refers to the amount of information contained in one random variable with another random variable and also a measure of the association between two variables. The formula of mutual information can be expressed as follows [17].

$$I(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y). \tag{12}$$

Mutual information (MI) is a commonly used information metric in information theory and is widely used in statistical language models. It measures the importance of a word to a category based on the occurrence of the word t. The calculation formula is shown as follows:

$$\mathrm{MI}(t) = \sum_{i=1}^{|C|} P(c_i)\log\frac{P(t|c_i)}{P(t)}, \tag{13}$$

where $P(t|c_i)$ denotes the conditional probability that feature t appears in category c_i, $P(c_i)$ denotes the probability that the text with category c_i appears in the document collection, and $P(t)$ is used to denote the probability that features t appears in the text.

From the formula, it can be seen that feature selection using mutual information prefers to select low-frequency feature words. That is, if two feature words have the same conditional probability $P(t|c_i)$, the feature word with fewer occurrences will receive a higher mutual information value than the feature word with more occurrences.

3. Validity Analysis of Online Review Texts

Online reviews sometimes have irrelevant content, simple copying and modification, and irrelevant content, which prevents tourists from obtaining valuable information from online reviews and also brings challenges to the operation of various online platforms [18, 19]. Therefore, it is very important to analyze the validity of online review texts. Validity analysis can be regarded as a two-classification problem of text, and it is divided into two categories: valid and invalid. That is, the subjectivity, randomness, irrelevant content, and suspicious copy and paste text in the sentence are classified into the "invalid" category, and the rest of the comment text belongs to the "valid" category [20]. The sample tested in this paper uses the web crawler technology Python 3.6 to crawl the review texts of scenic spots and hotels in China's major travel websites, especially choosing the online travel giant Trip as the main source of sample data.

The main steps about the validity analysis of online review texts are described below.

3.1. Data Preprocessing. To analyze the validity of the review content, it is necessary to preprocess the review content of scenic spots and hotels. The following two steps are included:

Step 1. Filter punctuation marks and stop words. Because there are many useless symbols in text information, many words have no practical meaning, such as some words. Therefore, the review data must be filtered first. The processing steps include filtering punctuation marks, special symbols, removing stop words, useless adverbs, etc. Here, we use a custom stop word dictionary to remove stop words more accurately and efficiently and to filter out a large number of words such as adverbs that have no practical meaning by word filtering.

Step 2. Word segmentation processing. Chinese word segmentation is not as simple as English word segmentation. There is no obvious distinguishing mark between words, and semantic and logical relationships are often taken into account. The effect of word segmentation directly affects information analysis and experimental results. At present, several common word segmentation tools are Jieba word segmentation, SnowNLP word segmentation tool, and HanLP word segmentation tool.

Jieba word segmentation is the most widely used word segmentation technology in China. The Jieba word segmentation has the following features: precise mode can segment sentences most accurately and is suitable for text analysis; the full mode can scan all the words that can be turned into words in the sentence, and the speed is very fast. SnowNLP word segmentation is a Python-based library. Its functions are relatively simple, and it is relatively easy to use. Jieba word segmentation is more suitable for text analysis than SnowNLP. Therefore, this paper uses Jieba word segmentation for Chinese word segmentation and word annotation.

3.2. Manual Annotation Validity Information. To build a machine learning model to analyze the effectiveness of reviews, the model must first be pretrained. So we made 1000 valid manual annotations in the reviews of scenic spots and hotels in advance. If there is useful information in the reviews of hotels and scenic spots, we regard it as a valid mark of "1." For some irrelevant or useless information, we regard it as an invalid mark of "0." The models are trained and tested accordingly through these manually annotated data.

Here, there are nearly 60,000 reviews of scenic spots and more than 25,000 reviews of hotels collected through crawler technology. Take the scenic spot as an example; we extract 1,000 items from them for manual annotation, that is, to add a column to the original data file. The column name is set to "valid," which is set to 1 when valid and 0 when invalid.

3.3. Extract Text Features. For the comment content of the segmented words, the numerical calculation of text feature is performed. In this paper, document frequency is used to calculate the value of text feature. Document frequency (DF) is an efficient feature selection algorithm that counts how many texts contain the word in the entire dataset. Its document frequency is counted for each feature in the training text set, and those features with a particularly low and particularly high document frequency are removed according to a preset threshold. DF is the simplest feature selection method, and this method has low computational complexity and can perform large-scale classification tasks. It is a common method for feature dimension reduction.

3.4. Construction of a Classification Model for the Validity of Online Review Texts Based on Naive Bayes Classifier. The classification method based on Naive Bayes classification algorithm is divided into two stages. The first stage is the training stage, building a classifier with a known set of instances. The known instance set used to build the classifier is called the training instance set, and each instance in the training instance set is used as a training instance. Since the class labels of the training instances are known, the construction of the classifier is a tutored learning process.

The second phase is the test phase, using constructed classifiers to classify unknown instances. The classifier generally needs to be evaluated before it can be used to predict. Only classifiers with the required classification accuracy can be used to classify test instances.

The classifier used here is the Naive Bayes classifier, and its characteristics mainly include incremental learning. Prior knowledge can determine the final probability of the hypothesis together with the observed instances and allow assumptions to make uncertain predictions. The classification of new instances can be predicted by multiple assumptions together weighted by their probabilities. Based on the above classification characteristics, this paper constructs a classification model for the validity of online review texts based on the Naive Bayes classifier. The specific process is using the Naive Bayesian method for supervised learning and training based on the 1,000 manually annotated scenic spot reviews. These trained models are then used to annotate all the online review texts. Then, for the annotated review file table of hotels or scenic spots, a column is added to the right: "validity," and it is marked as "valid" or "invalid" according to the output of the trained model. Finally, the validity analysis table of tourism online review texts is obtained, as shown in Table 1.

4. Comprehensive Evaluation of Scenic Spots and Hotels Based on Text Classification Technology and Sentiment Analysis

After removing the invalid text, the effective online review texts of scenic spots and hotels are obtained here. We use text classification techniques to classify online review texts into appropriate categories. According to the five aspects of scenic spots and hotels that tourists focus on, service, location, facilities, hygiene, and cost, a comprehensive evaluation is carried out, and an evaluation model is constructed according to mean squared error (MSE) combined with sentiment analysis. Mean squared error is a measure that reflects the degree of difference between the estimator and the estimated. Let t be an estimator of the population parameter θ determined from the subsample. The mathematical expectation of $(\theta - t)^2$ is called the mean squared error of the estimator t. $\text{MSE} = \sigma^2 + b^2$, where σ^2 and b^2 are the variance and bias of t, respectively.

Therefore, this paper uses the mean squared error to build the evaluation model according to the above data analysis. The specific steps are as follows.

(1) This paper uses web crawler technology to crawl corpus in five aspects of service, location, facility, hygiene, and cost from major tourism websites and performs part of manual annotation and model training

First, we look for the text classification corpus for training, where the text files containing about 5,000 hotel online reviews are collected. According to the five aspects of service, location, facilities, hygiene, and cost, manual annotation is carried out one by one, and more than 1,500 online comment texts are marked and classified into the abovementioned five categories.

(2) The paper splits the review texts of hotels or scenic spots into single sentences and then classifies the texts according to service, location, facilities, hygiene, and cost

According to the data characteristics of online review texts in the tourism field, two suitable text classification methods are selected for comparative testing. The best text classification method is selected according to the processing speed and classification evaluation indicators.

Two text classification methods are used here: one is Naive Bayesian method, and the other is linear SVM. The sample data are tested separately, and the experimental

Table 1: Validity analysis table of tourism online review texts.

Hotel name	Comments	Comment validity
H01	The hotel is suitable for family travel.	1
H01	Upgraded the room, late checkout, and it is great.	1
H01	I have come to Guangzhou every year, and I will stay in a ** hotel, because the location is good and the price/performance ratio of the hotel is also good.	1
H02	The hotel is very good.	0
H02	Super 5 stars.	0
H02	The hotel is clean and hygienic; the service is very good!	1
H03	The hotel is a traditional standard five-star hotel. The only regret is that the bathroom only has a bathtub and no shower.	1
H03	Hotel facilities are a bit old, but the location is really good.	1
H03	Very good location and convenient travel.	1
H04	Good location, clean, and tidy room.	1
H04	The hotel is in a great location! 100 m out of the metro entrance. And the service is fantastic! The hotel management is improving!	1
H04	The hotel is a good choice.	0
H05	It is very close to Guangzhou East Railway Station, convenient for picking up and dropping off guests, the price is reasonable and acceptable, the breakfast is good and filling, and the variety is complete, suitable for young and old women and children to choose their own food.	1
H05	Jianguo is always a very good choice for business travelers who want to sleep in a serviced area in the Central district of Tianhe.	0
H05	High cost performance, rich breakfast varieties, many kinds of bread, close to the subway.	1
H06	It is very close to Guangzhou East Railway Station and subway station, and it is very convenient to eat nearby. The price/performance ratio is quite good. Recommended to stay.	1
H06	The room is very good; the service is very good; the front desk is very good.	1
H06	Very good for many times and cost-effective.	1
H07	The hotel is relatively close to Guangzhou East Railway Station, just next to it.	1
H07	Hygiene is very good.	1
H07	Good location and good service.	1
H07	The hotel is very good; the service is very good.	1

results show that the linear SVM method is faster in terms of data processing speed, as shown in Figure 1.

In this paper, the indicator test of the text classification model is carried out. The model evaluation indicators of the classification algorithm are often measured by the confusion matrix, as shown in Table 2. The four data obtained from the confusion matrix are extended by calculation to obtain four secondary metrics: accuracy, precision, recall, and F1-score, which are the core metrics for evaluating the classification model [21].

Accuracy of a classification model (accuracy) represents the ratio of samples correctly predicted by the model to all samples, and in general, the higher the accuracy, the better the classifier is accordingly. This is shown in the following equation.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (14)$$

The precision of the classification model (precision) is defined as the percentage of samples with true positive class among all samples predicted to be positive class, and the formula is as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (15)$$

The recall (recall) of a classification model is defined as the percentage of samples with true positive classes that are correctly predicted, and the formula is as follows:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (16)$$

F1-score is the summed mean of precision and recall, see equation (17), which combines the results of precision and recall and is closer to the smaller of the two, so when precision and recall are close, F1 is the largest. A higher value of F1 indicates a better model prediction.

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (17)$$

Here, the test is carried out on the sample data. First, the

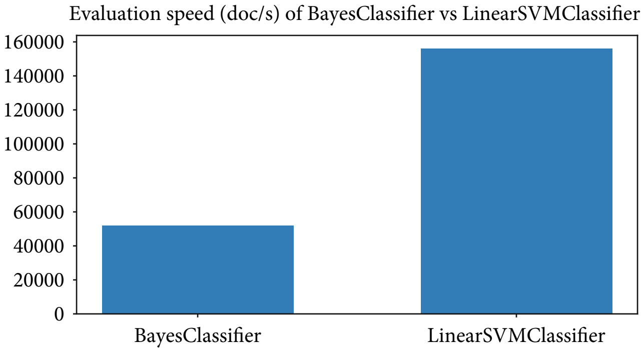

Figure 1: Comparison of the processing data speed between the two classification methods.

Table 2: Confusion matrix.

Predict True	0	1
0	TP	FN
1	FP	TN

Naive Bayes method is used. According to the above four classification indicators, the test results in five aspects of service, location, facilities, hygiene, and cost are shown in Figure 2.

In order to compare the classification effect, this paper uses the linear support vector machine to test the five aspects (service, location, facilities, hygiene, cost) on four classification indicators. The test results are shown in Figure 3. The experimental results show that for the online review texts tested in this paper, the linear support vector machine has better classification effect than Naive Bayes method on the four classification indicators by synthesizing the five aspects of tourists' concerns. Finally, linear support vector machine is selected as the classification model of online review text.

(3) A single evaluation is made on the 5 aspects of each hotel or scenic spot (service, location, facilities, hygiene, cost); combined with text sentiment analysis, this paper adopts a 5-point system for comprehensive scoring

On the basis of text classification, all online review texts of hotels can be classified into the above 5 categories (service, location, facilities, hygiene, cost). After the classification is done using the classifier, a confusion matrix with specific numerical values can be shown by Figure 4.

For example, in the "service" category, a certain online review under this category is scored according to objective criteria. Sentiment analysis in natural language processing is used here. Sentiment analysis is a classification technique based on natural language processing. The main purpose of sentiment analysis methods is to determine whether a review is positive or negative. Therefore, this paper uses sentiment score indicators to quantify online review texts. Sentiment analysis generally sets the sentiment of the text to a value between (0,1). 0.5 represents neutrality, a score closer to 0 represents a negative emotion, and a score closer to 1 represents a positive emotion. Since the score is based on a 5-point system, the result of sentiment analysis is multiplied by 5 to get a score between (0,5). After all the online review texts of the hotel under the category of "service" are scored on the above 5-point system and averaged, the "service" score of the hotel can be obtained.

In the same way, the other four aspects of the hotel can be scored. In addition, the same method can be used to score the five aspects of the scenic spot.

(4) The evaluations of all hotels or scenic spots are normalized and comprehensively evaluated, and a score between 1 and 5 is obtained. The experimental results are shown in Figure 5

The experimental results in Figure 5 show that Hotel01's scores in five aspects are 4.5, 4.6, 4.3, 4.5, and 4.7, and the comprehensive score is 4.5; Hotel03's scores in five aspects are 4.3, 4.6, 4.1, 4.5, and 4.4, and the comprehensive score is 4.4; Hotel05's scores in five aspects are 4.4, 4.6, 3.9, 4.5, and 4.5, and the comprehensive score is 4.4. Both Hotel02 and Hotel04 have the same comprehensive score of 4.2. Therefore, this paper uses the sentiment analysis method in natural language processing to comprehensively score the five services of the hotel, and the hotel with the highest score can be selected as Hotel01, which can provide decision-making basis for hotels and scenic spots to improve service quality.

5. Mining and Analysis of Tourism Hot Words Based on Natural Language Processing

Hot words are the most direct form to reflect tourists' impression of scenic spots and hotels. Generally speaking, whether it is positive or negative, tourists will always have some representative words in their comments on scenic spots and hotels. These words are often also high-frequency words, which are obtained by statistical methods after excluding irrelevant stop words. The hot words need to be obtained through natural language processing. This paper obtains the hot words in the online review text based on the named entity recognition method.

Named entity recognition is a common basic task in natural language processing, and it is also an important component of tasks such as information extraction, question answering, syntactic analysis, and machine translation. Common types of entities are person names, place names, institution names, times, dates, money, and so on. Different tools have some subtle differences in the distinction between entity types. This article is concerned with a few types of place names, such as hotels or scenic spots [22].

In this paper, the idea of mining hot words from online review texts is as follows. Firstly, according to the above method, the scenic spots and hotels are comprehensively evaluated and scored, and the thresholds are set into three levels: high, medium, and low, and the top 10 of

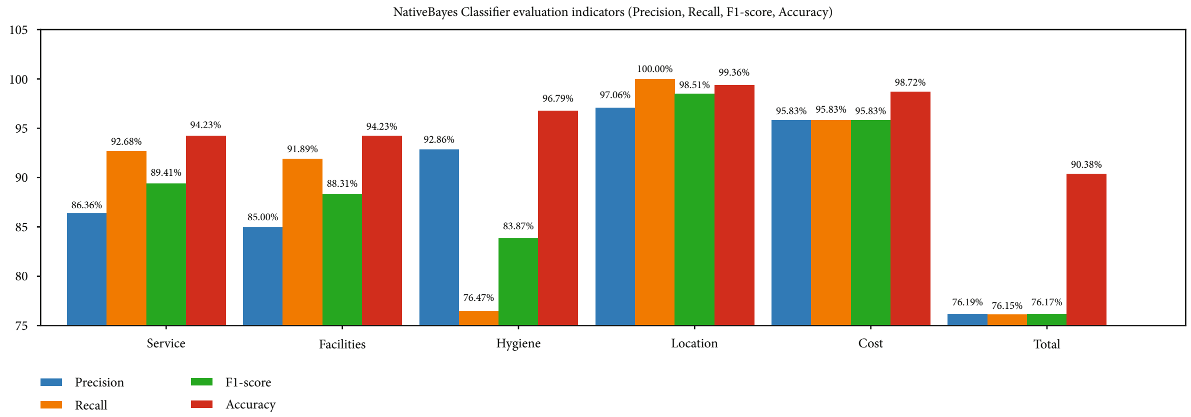

Figure 2: Results of the classification indicator test based on Naive Bayes.

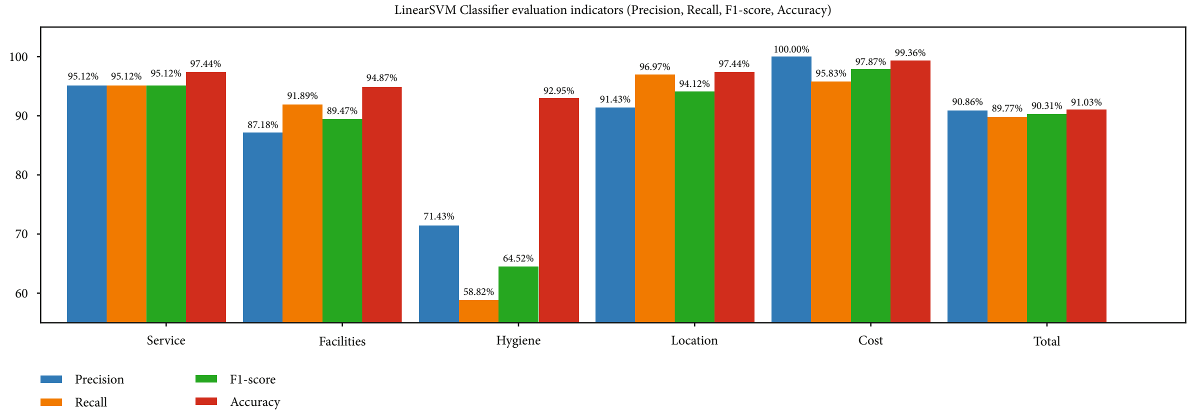

Figure 3: Results of the classification indicator test based on linear support vector machine.

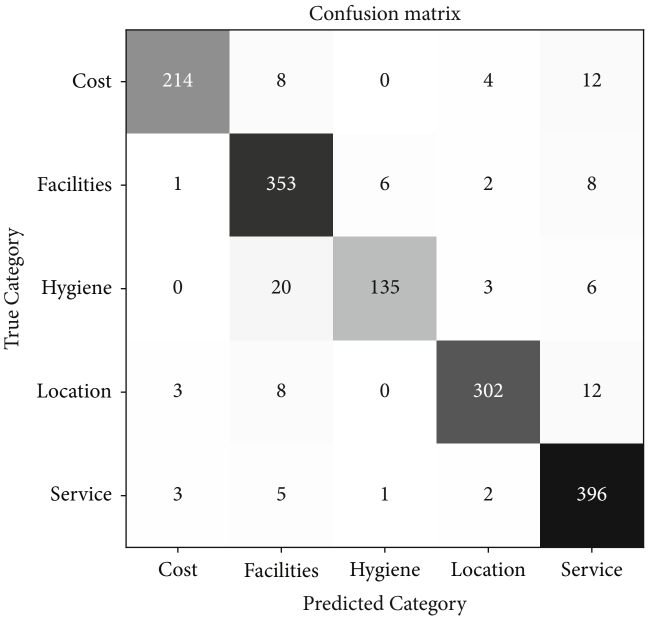

Figure 4: Confusion matrix with specific values.

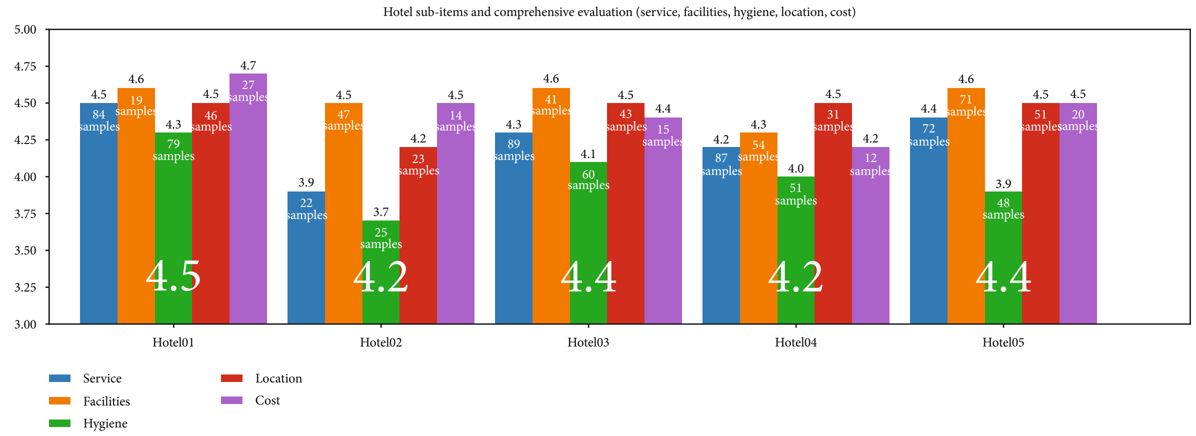

Figure 5: Results of comprehensive hotel scoring based on natural language processing and sentiment analysis.

each level are taken as the online comment text for mining hot words. These online comment texts are then named entity recognition, and the obtained named entities (noun) are stored in the word segmentation dictionary. Then, the comment text is segmented based on the named entity as a word segmentation dictionary, and after filtering the part of speech, high-frequency words are counted, and finally, hot words are obtained. Based on this idea, a flow chart of hot word acquisition is formed, as shown in Figure 6.

(1) *Corpus Preprocessing*. Here, the Python module of pandas is used to read the corpus according to the scenic spots, and the Python module of the natural language processing tool SnowNLP is used to segment superlong text into short sentences. Since the NLP tool (HanLP 2.x) used in the following steps cannot handle superlong text, during preprocessing, these superlong sentences need to be segmented into short sentences to lay the foundation for subsequent processing

(2) *Named Entity Recognition*. Using the NER function in the NLP tool, the words with named entities such as scenic spots and facilities are extracted from each sentence and written to the word segmentation dictionary

After comprehensive testing and comparison, the NER function provided in HanLP 2.x version is adopted here. The reason is that this version uses a large-scale Chinese corpus pretrained deep learning model and provides three mature methods: ner/msra, ner/pku, and ner/ontonotes. The effect of the test here is relatively good, and all the scenic spots can be found. Here, the named entities obtained by the NER methods are merged and deduplicated as the actual NER results.

(3) *Word Segmentation*. In the case of using a word segmentation dictionary of NER results, the word segmentation tool is Jieba. This paper uses this tool to perform word segmentation on all the review corpus of a scenic spot or hotel. In order to ensure the effect, only the noun is retained in the result after word segmentation

(4) *Statistics of Word Frequency*. For the list of word segmentation results, the word frequency of all words is counted. The 20 words with the highest frequency are found and written to the file in the required format

The hot words generated by 16 scenic spots such as A01-A16 are listed in Table 3. From the experimental results in Table 3, it can be seen that hot words can effectively display the intuitive impression of scenic spots to tourists.

6. Analysis Characteristic Services of Scenic Spots and Hotels Based on New Word Discovery

In order to attract tourists and enhance their competitive advantage, scenic spots need to find their own characteristics. This paper excavates their respective characteristics and highlights from the online review texts of scenic spots and hotels. Combined with the comprehensive evaluation results of scenic spots and hotels, after the appropriate threshold value is determined, the scenic spots and hotels are divided into three levels: high, medium, and low. And for the top three of each level, this paper applies the new word discovery method to find the characteristics of scenic spots and hotels.

New word discovery is also called "new word recognition" or "unregistered word recognition." The new words here are words that are newly generated with the development of the times, for example, "cloud era," "big data," letter words "yyds," and "xdm." Alternatively, words that do not exist in the dictionary may also be called new words or unregistered words [23].

New word can be judged as "new" from three aspects:

(1) Degree of solidification refers to the degree of closeness between words in a field. For example, words

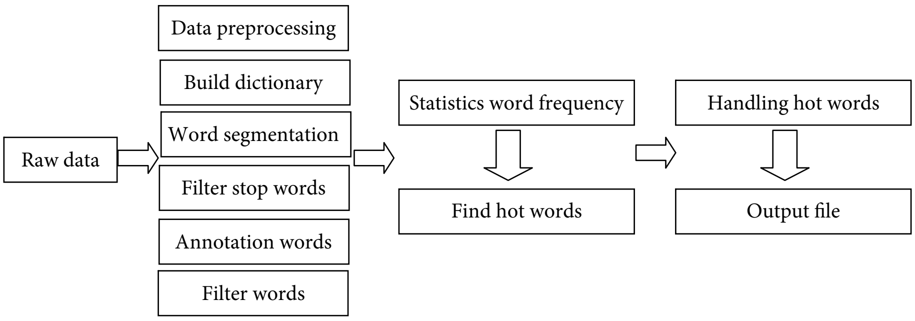

Figure 6: Process flow chart of mining hot words.

Table 3: Hot word list of online review text based on natural language processing.

Scenic area name	Hot words
A01	Animal, zoo, circus, project, scenic, queue, roller coaster, children, ticket, wonderful, train, park, feel, place, tour, and advice
A02	Fireworks, children, queue, animals, project, time, roller coaster, play, place, ocean, ocean kingdom, wonderful, scenic, park, facilities, and feel
A03	Project, tea valley, scenery, queue, place, scenic, Grand Canyon, attractions, feeling, time, roller coaster, small fire, tour, facilities, and environment
A04	Project, queue, ID, roller coaster, time, facilities, feel, place, a little, hours, night, haunted house, snowy, Halloween, and advice
A05	Attractions, feel, place, Shenzhen, architecture, world, guide, tickets, project, night, scenic, advice, tour, queue, scenery, and landscape
A06	Guangzhou, night view, Ferris wheel, queue, small barbarian waist, scenery, landscape, landmark, Pearl River, building, boarding tower, and elevator
A07	Attractions, place, folk village, scenic spot, China, feel, scenery, wonderful, program, ethnic, landscape, Shenzhen, tour, culture, night, and advice
A08	Animal, zoo, child, kids, tickets, feel, tiger, scenic, place, wonderful, time, small children, project, show, kind, and environment
A09	Scenic spot, attraction, scenery, view, place, climbing, time, feeling, tour, yang yuan mountain, mountain, and advice
A10	Game, project, animals, children, motorized, children, play, zoo, place, play, facilities, queue, a little, children, tickets, and feel
A11	Place, landscape, scenery, project, wedding photos, play, children, scenic, attractions, games, very beautiful, environment, wedding, and queue
A12	Fish, place, environment, scenery, attractions, play, garden, landscape, a little, park, Lingnan, feel, characteristics, architecture, scenic, and goldfish
A13	Hot spring, front desk, environment, service attitude, enthusiasm, waiter, soak hot spring, hot spring pool, fruit, attitude, pool, water, and staff
A14	Underground river, scenery, attractions, scenic spot, cave, queue, guide, cave, place, feel, scenery, tour, stalactites, very beautiful, and book tickets
A15	Animal, place, mermaid, feel, beluga, hour, penguin, little kids, fish, ocean, polar bear, show, and time
A16	Scenic spots, scenery, places, Guanyin, scenery, Nanhai, Foshan, environment, up the mountain, climbing, famous mountain, Xiqiao, Guangdong, and walking

such as “research” and “water cup” have a relatively high degree of solidification, while words such as “Haier” and “Gree” have a relatively low degree of solidification

(2) Degree of freedom refers to the degree to which a field can be used freely. For example, “Jujue” has the same degree of solidification as “Jejue,” but the freedom degree of “Jejue” is far less than that of “Jejue”

(3) The IDF (inverse document frequency) of new words is called the inverse document frequency of new words. If a word appears a high number of times in an article, it is most likely a new word. But if the word also appears a high number of times in the entire text corpus, then it may be a common word, not a new word

There are generally two approaches to new word discovery: rule-based new word discovery and statistics-based new

Table 4: The characteristics of scenic spots obtained by the new word discovery.

Scenic spots name	Scenic features
A01	Worlds of fun and online booking
A02	Motorized games, whale shark house, and penguin hotel
A03	Amusement facilities, mesa, tragen, motorized games, and online booking
A04	Direct swipe, Maya water, snowy eagles, rides, Snowy Mountain Flying Dragon, and water world
A05	Philadelphia tower, continental area, reduced version, itinerary, miniature view, and famous buildings
A06	Jumper, observation deck, landmark building, extreme skyscraper, landmark, Zhujiang New Town, world no. 1, and Haixinsha
A07	Miniature landscape and oriental neon
A08	Direct swipe, distance contact, and water park
A09	Changlao peak, Yang yuan stone, yin yuan stone, Danxia on water, Hosomei Zhai, and buy on site
A10	Motorized games, pick up tickets, suitable for a family, clip the doll, and clip the doll
A11	Wedding photography, with children, motor games, and wedding photography
A12	Pick up tickets, suitable for a family, and ancient fragrance
A13	Pick up tickets, no borders, and fruit drinks

Table 5: The characteristics of hotels obtained by the new word discovery.

Hotel name	The characteristics
H01	Near the metro, reception, next to East Station, old five star, free upgrade, full facilities, high train station, and clean and tidy
H02	Shennan avenue, convenient transportation, and full facilities
H03	Chang long, impressed, and great location
H04	Suitable for children and convenient transportation
H05	Free upgrade, easy access, platinum 5 stars, and clean and hygienic
H06	Suitable for strip, De Beer Plaza, and free upgrade
H07	Yutang spring, old five star, free upgrade, Sha Mian Park, hometown water, Servcorp house, toiletries, White Swan Lake, and redecoration
H08	Clean and tidy, fully equipped, free parking, front desk reception, close to West Lake, and great location
H09	Suitable for children, breakfast is plentiful, and clean and tidy
H10	Sea World, free upgrade, clean and tidy, check in, and decoration style
H11	Suitable for the strip, fairview, artificial beach, and breakfast buffet
H12	Apartment underground station, free upgrade, affordable, free parking, at the underground entrance, sound proofing, with character, and eating and shopping
H13	Changlong, lovers' road, breakfast is plentiful, and fully equipped

word discovery [24]. New word discovery based on rule can be performed by constructing a template for new word matching, and the obtained results have a relatively high accuracy rate. New word discovery based on statistics is to identify new words by counting the word frequencies in the corpus. This method is more portable and flexible and requires a certain model for training. In this paper, an algorithm based on mutual information and left-right entropy is used to discover new words. The specific steps are as follows:

(1) Stop word processing

The text that needs to be processed often contains many meaningless words. This paper replaces these texts or removes stop words to get cleaner texts.

(2) Use three thresholds to judge new words

(i) *Minimum Mutual Information*. The greater the mutual information, the higher the correlation. This paper uses the n-gram software for word segmentation and then calculates mutual information for these words. If it is below the threshold, it means that it cannot be a word

(ii) *Minimum Entropy*. The larger the entropy, the more abundant the neighboring words. This paper calculates the minimum value of left entropy and right entropy. If the minimum value is lower than the threshold value, it means that the word cannot be formed

(iii) *Minimum Number of Occurrences*. If the number of occurrences of a word is less than the set minimum number of occurrences, it is filtered out

(3) Mining of characteristic in scenic spots and hotels

In this paper, the algorithm based on mutual information and left-right entropy is used to discover new words. At the same time, this algorithm also integrates the Python module of SmoothNLP. The implementation of SmoothNLP is different from the work of mutual information and left-right entropy. For the same text, the new words discovered by the two methods are not consistent. Therefore, we calculate the two methods together and obtain the common vocabulary of the two methods. In addition, natural language processing also provides an option "whether to extract only words that are not in the dictionary." Thereby, the search scope can be expanded to find more unregistered new words. Therefore, this pattern is also incorporated into the method of new word discovery. Here, these three methods are combined to discover new words in the online review texts by adjusting the parameters. This new word discovery method considers both the breadth and novelty of new words and the acceptability of new words, which can reflect the characteristics of scenic spots or hotels. Finally, the test is carried out on the online review text with high comprehensive evaluation, and the experimental results show: Table 4 shows the characteristics of 13 scenic spots obtained by the new word discovery algorithm. Table 5 shows the characteristics of the 13 hotels obtained by the new word discovery method.

7. Conclusion

At present, it is a research hotspot in the field of tourism management to provide directions for the development of tourism by the mining of online review texts. This paper first analyzes the research status and finds that the current application of online review texts is not systematic, and there is a lack of a comprehensive scoring mechanism for the service quality of scenic spots and hotels. Then, this paper proposes a novel mining and application of tourism online review text based on natural language processing and text classification technology. A series of new methods are proposed here.

The first is to remove the invalid online review text and keep the valid text. The research focus of this paper is to propose a comprehensive evaluation of scenic spots and hotels based on text classification technology and sentiment analysis methods. This evaluation system can establish evaluation indicators for comprehensive scoring and select top-ranked scenic spots and hotels. Then, for the online review texts of high-quality scenic spots, this paper proposes to use natural language processing to mine hot words. The obtained hot words can intuitively reflect the impression of the scenic spot to tourists. Then, new words are discovered based on mutual information and left-right entropy methods. The service characteristics of high-quality scenic spots and hotels are excavated from new words. Finally, the above series of methods are tested on the massive tourism online review texts. The experimental results show that the new comprehensive evaluation method proposed in this paper can effectively select high-quality scenic spots and hotels. Hot words and new words can be efficiently mined from relevant online review texts. These methods can feedback tourism impressions from various aspects and levels and provide important basis for the development of tourism. Due to the limitations of the selected online review texts, the comprehensive evaluation method in this paper has certain regional characteristics and is not suitable for all tourist locations. There are two aspects of future research work. The first is to expand the crawling scope of online review texts as much as possible and establish more comprehensive evaluation indicators. The second is to consider how to use the discovered hot words and new words to provide personalized intelligent services for tourists and tourism practitioners.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Funds of China (61272015) and the 2022 Henan Province Science and Technology Research Project: "Construction and Application of Intelligent Ontology in Internet of Things Based on Semantic Concept Model" (222102210316).

References

[1] W. Chen, Z. Xu, X. Zheng, Q. Yu, and Y. Luo, "Research on sentiment classification of online travel review text," *Applied Sciences*, vol. 10, no. 15, p. 5275, 2020.

[2] X. Li, R. Law, G. Xie, and S. Wang, "Review of tourism forecasting research with Internet data," *Tourism Management*, vol. 83, no. 4, article 104245, 2021.

[3] W. Yuan, X. Xin, and F. Xuegang, "Research on tourist' percieved image of ancient town using web text mining methods: a case study of Zhujiajiao," *Tourism Science*, vol. 27, no. 5, pp. 86–94, 2013.

[4] C. A. I. Yi, Y. A. N. G. Yang, and Y. Hongmei, "Research on audience's perception of tourism brand of Guiyang based on the text mining of ROST," *Journal of Chongqing Normal University(Natural Science)*, vol. 32, no. 1, pp. 126–134, 2015.

[5] W. Yaobin, Y. Ling, and S. Chuanling, "Comparative research on travel sharing of typical travel website based on text mining-taking Gansu Province as an example,"

Resource Development & Market, vol. 33, no. 1, pp. 100–104, 2017.

[6] L. Ping, C. Tian, and W. Fuyuan, "Urban tourism community image perception and differentiation based on online comments: a case study of Beijing," *Geographical Research*, vol. 36, no. 6, pp. 1106–1122, 2017.

[7] L. I. N. Xuanmiao, "A comparative study of tourists' opinions and suggestions on Qinbeifang of Guangxi based on text mining," *Market Forum*, vol. 7, pp. 69–74, 2017.

[8] H. Ming, S. Jianjun, and C. Ying, "Text classification based on Naive Bayes: a review," *Information Science*, vol. 34, no. 7, pp. 147–154, 2016.

[9] L. Lv, Z. Wu, L. Zhang, B. B. Gupta, and Z. Tian, "An edge-AI based forecasting approach for improving smart microgrid efficiency," *IEEE Transactions on Industrial Informatics*, vol. 3, pp. 1–1, 2022.

[10] L. Lv, Z. Wu, J. Zhang, Z. Tan, L. Zhang, and Z. Tian, "A VMD and LSTM based hybrid model of load forecasting for power grid security," *IEEE Transactions on Industrial Informatics*, vol. 11, p. 1-1, 2021.

[11] G. U. A. N. Lei and S. U. N. Tao, "An efficient parallel and distributed solution to nonconvex penalized linear SVMs," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 4, pp. 587–603, 2020.

[12] L. Zhang, Z. Huang, W. Liu, Z. Guo, and Z. Zhang, "Weather radar echo prediction method based on convolution neural network and long short-term memory networks for sustainable e-agriculture," *Journal of Cleaner Production*, vol. 298, article 126776, 2021.

[13] Y. Lei Zhang, Q. G. Huo, Y. Ma, Q. Liu, and W. Ouyang, "A privacy protection scheme for IoT big data based on time and frequency limitation," *Wireless Communications and Mobile Computing*, vol. 2021, 10 pages, 2021.

[14] Y. Ye, Q. Wu, Y. Li, K. P. Chow, L. C. Hui, and S. M. Yiu, "Unknown Chinese word extraction based on variety of overlapping strings," *Information Processing & Management*, vol. 49, no. 2, pp. 497–512, 2013.

[15] L. Zhang, C. Xu, Y. Gao, Y. Han, X. du, and Z. Tian, "Improved Dota2 lineup recommendation model based on a bidirectional LSTM," *Tsinghua Science and Technology*, vol. 25, no. 6, pp. 712–720, 2020.

[16] L. Zhang, S. Tang, and L. Lv, "An finite iterative algorithm for sloving periodic Sylvester bimatrix equations," *Journal of the Franklin Institute*, vol. 357, no. 15, pp. 10757–10772, 2020.

[17] L. Lv, J. Chen, Z. Zhang, B. Wang, and L. Zhang, "A numerical solution of a class of periodic coupled matrix equations," *Journal of the Franklin Institute*, vol. 358, no. 3, pp. 2039–2059, 2021.

[18] C. Nan, Z. Jian, and W. Juqing, "Not only rating, but also text content will produce an influence on review helpfulness," *Luojia Management Review*, vol. 1, pp. 15–26, 2014.

[19] D. Jianyong, G. Huijuan, and Z. Mei, "Research on the method of network review extraction," *Journal Of North China University*, vol. 27, no. 1, pp. 7–12, 2015.

[20] L. Lv, S. Tang, and L. Zhang, "Parametric solutions to generalized periodic Sylvester bimatrix equations," *Journal of the Franklin Institute*, vol. 357, no. 6, pp. 3601–3621, 2020.

[21] D. Brzezinski and J. Stefanowski, "Prequential AUC: properties of the area under the ROC curve for data streams with concept drift," *Knowledge and Information Systems*, vol. 52, no. 2, pp. 531–562, 2017.

[22] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A review of text corpus-based tourism big data mining," *Applied Sciences*, vol. 9, no. 16, p. 3300, 2019.

[23] A. Usai, M. Pironti, M. Mital, and C. Aouina Mejri, "Knowledge discovery out of text data: a systematic review via text mining," *Journal of Knowledge Management*, vol. 22, no. 7, pp. 1471–1488, 2018.

[24] L. Weitong, L. Peiyu, and L. Wenfeng, "New word discovery algorithm based on mutual information and branch entropy," *Application Research of Computers*, vol. 36, no. 5, pp. 1294–1296, 2019.