Hindawi
Journal of Sensors
Volume 2022, Article ID 3365498, 9 pages
https://doi.org/10.1155/2022/3365498

Research Article

Emotional Calculation Method of Rural Tourist Based on Improved SPCA-LSTM Algorithm

Xi Chen [1,2,3]

[1] *College of Geography and Tourism, Harbin University, Heilongjiang, Harbin 150086, China*
[2] *Heilongjiang Province Key Laboratory of Cold Region Wetland Ecology and Environment Research, Heilongjiang, Harbin 150086, China*
[3] *Harbin Institution of Wetland Research, Heilongjiang, Harbin 150086, China*

Correspondence should be addressed to Xi Chen; chenxi005253@126.com

Received 4 January 2022; Revised 24 January 2022; Accepted 5 February 2022; Published 23 February 2022

Academic Editor: Wen Zeng

New technologies such as big data and cloud computing provide new means and tools for rural development. The big rural tourism, with its convenience, quickness, and low threshold, presents great convenience for tourists' emotional calculation and has become one of the main sources of tourism big data. Under the guidance of big data theory and emotion theory, this paper proposes an emotional calculation method of rural tourists based on improved SPCA-LSTM algorithm, taking big text data as data source. Firstly, the improved TF-IDF algorithm is designed to highlight the importance of feature items, and the word vector trained by word2vec model is applied to represent the rural tourism data text. Then, a weighted sparse PCA (RSPCA) is constructed to reduce the dimension of massive word vector features. RSPCA introduces the weighted l_1 optimization framework and LASSO regression model into the mathematical model of PCA algorithm and establishes a new data dimension reduction model. Thereupon, the long-term and short-term memory convolution network with attention mechanism is employed to extract text features. Finally, the feature vector is utilized to calculate the rural tourist's emotion by softmax function. The experimental results indicate that the improved SPCA-LSTM algorithm, whose performance index is better than other existing algorithms, is effective in calculating tourists' emotions. Also, it is more suitable for the research of tourist sentiment calculation in the era of big data.

1. Introduction

Rural revitalization is an important part of urban economic development and a beneficial assist for urban-rural integration. Rural tourism is a relatively efficient way that China has explored during poverty alleviation for more than 30 years, and a further way to overcome the problem of rural revitalization during the "14th five-year plan" [1]. New technologies, such as artificial intelligence, 5G, and big data, bring infinite possibilities to China's economic and social development, provide new means and tools for rural development, and are powerful support for rural revitalization in the new era. Not only does digital technology empowering rural revitalization change the traditional production mode and lifestyle, but also subtly changes farmers' practice and thinking mode. Promoting rural revitalization and industrial prosperity is the endogenous power support. Rural characteristics and the law of information development should be respected. For villages with rich local characteristic resources and relatively complete information facilities, they can rely on "Internet +" to dig local characteristic cultural resources and do a good job of digital "cultural tourism" and "agricultural tourism" combined with articles to build Internet characteristic tourism villages [2].

With the rise of new media and social networks and the advent of the era of big data, various platforms have generated hundreds of millions of rural tourism-related data. The past data showcase that since 2009, the amount of data related to rural tourism has increased geometrically year by year. Due to the convenience of social transportation and the development of social networks, the data with "rural tourism" as the key word is widely spread in rural areas. People's demand for "rural tourism" is increasing day by day. The data of "rural tourism" can be analyzed and mined through various information release platforms [3, 4].

Rural tourism reception will increase steadily. Data show that from January to August 2020, the number of leisure agriculture and rural tourism in China decreased by 60.9% to 1.207 billion. The epidemic situation in COVID-19 has pushed the "pause button" for leisure agriculture and rural tourism. However, after the domestic epidemic was cleared, all localities restarted the rural leisure tourism market in an orderly manner. People's enthusiasm for rural scenery and fresh air is high. With the gradual restoration of production and living order, the suppressed demand of urban and rural residents will continue to be released. The countryside with beautiful scenery and beautiful ecology is more attractive than ever before [5].

The rapid development of information technology provides an efficient and rapid way for rural tourist information acquisition [6]. The main reason for the popularity of rural tourism is that it provides tourists with multilevel experiences, and this information can be obtained through the classification of rural tourism big data driven by the rural revitalization strategy. The trend of rural tourism can be analyzed according to different types, different times, different spaces, and different groups. On various data platforms, search and collect data for specific keywords, and deeply analyze these data to get the corresponding results. The long-term data are verified and used as prediction support. Data mining on the future rural tourism market can make suggestions for future rural tourism [7].

Rural tourism is an activity of interpersonal communication and emotional exchange, and emotion runs through the whole process of tourism activities. "Emotion is a person's attitude experience about whether objective things can meet his own needs." Tourist emotion refers to the pleasure, excitement, sadness, anger, regret, and other emotional experiences generated by tourists in tourism activities due to the influence of personal factors or external environment on whether tourism activities meet individual basic needs and social needs and presents diversity and variability with the progress of tourism [8]. These emotions not only constitute an important tourist experience but also exert an important influence on tourist motivation, satisfaction, behavioral intention, and interpersonal interaction. In tourism, tourists get information and share their travel experiences through online platforms and social media. The text, image, audio, and video released by it become the main data source of tourism big data [9]. Among them, the texts content with its convenient, simple, intuitive, fast, and low threshold for tourists to express emotions and information exchange to provide convenience; in the tourism, big data occupies an increasingly important position. The mining of text data can provide decision support for tourism planning and marketing, making sentiment analysis in tourism big data a hot issue in tourism research [10].

With the deepening of research, emotion analysis, which is to effectively analyze and mine information and identify emotional tendencies, becomes more sophisticated. Also, there are researches on emotion calculator, emotion summary, product attribute mining, and so on. In recent years, with the development of big data, there are many emotion analysis models and software at home and abroad, which provide strong support for emotion research. There are three methods of text sentiment analysis: dictionary-based method, machine learning method, and deep learning method [11]. The dictionary-based approach is mainly through the development of a set of emotional dictionaries and rules. Then, sentiment value is calculated by sentence breaking, analysis, and dictionary matching. Finally, the emotional value is used as the basis to judge the emotional tendency of the text. Although the method based on emotion dictionary has high accuracy, it has low recall rate. Another problem of the dictionary-based approach is the relatively high cost of dictionary construction [12]. The machine learning-based approach transforms the problem of text sentiment analysis into a supervised classification problem. Annotate the training text. Supervised machine learning is then carried out [13]. Finally, the test data is used to predict the results through the model. The method based on machine learning generally includes two steps: text preprocessing and the selection of classification algorithm. Methods based on deep learning use different artificial neural network models to map text big data into vector space to obtain numerical representation of words. The numerical expressions are then used as input to the deep learning model. The optimal model is obtained through training and parameter optimization. The process is much the same as that of emotion computing based on machine learning. However, the selection of feature extraction and classification model for text vectorization is different from machine learning [14, 15].

This paper proposes a rural tourist emotional calculation method based on the improved SPCA-LSTM algorithm according to the above research. The algorithm improves the traditional TF-IDF algorithm by considering the distribution and location information of feature items within and between classes and combines it with Word2vec word vector to express the text. Then, reweighted SPCA algorithm is used to reduce the dimensionality of the text word vector, so that the original data features of the sample can be retained as much as possible while reducing the dimensionality. Finally, long-short-term memory (LSTM) and attention mechanism network are used to realize the emotional computing of rural tourists.

2. Tourist Emotion Calculation Method in This Paper

2.1. Text Representation and Feature Selection (Word Vector Representation). In view of the shortcomings of traditional TF-IDF in the field of text classification, this paper improves the traditional TF-IDF algorithm to better reflect the importance of feature items in the text.

(1) Introducing intraclass factors to describe the distribution relationship of feature items

The intraclass factor inter C_i is introduced to judge the uniformity of the distribution of feature words in the intraclass documents, that the intraclass factor value is small means that the feature items are unevenly distributed in the class and may be distributed only in a few documents in the class, with weak classification ability. On the contrary, the feature item has strong classification ability. The calculation formula of in-class factor of characteristic term is shown in

$$\text{inter}\, C_i = \frac{1}{S_{ij}}, \tag{1}$$

where j represents the category. S_{ij} represents the standard deviation of feature item i in category j, which reflects whether feature items are evenly distributed in the category. The smaller the standard deviation, the more uniform the distribution, the stronger the classification ability of this feature word. The calculation of S_{ij} is shown in formula (2).

$$S_{ij} = \sqrt{\frac{\sum_{p=1}^{K} \left(tf_{ip} - \bar{tf}_{ij}\right)^2}{K}}, \tag{2}$$

where K is the total number of documents in category j. tf_{ip} indicates the number of times that document p contains feature i. $\bar{tf}_{ij}$ represents the mean number of times of all documents of feature word i in category j. $\bar{tf}_{ij} = 1/K\sum_{k=1}^{K} tf_{ik}$.

(2) Improve the discrimination between classes

In order to improve the weight of feature items with high classification, m/m_i is used to express the importance of feature item i in category. N/n_i indicates the importance of feature items in all documents, where N represents the total number of documents in the corpus. n_i indicates the number of documents in which the feature word i appears in the corpus. m represents the total number of categories. m_i indicates the number of categories containing feature items. If the m/m_i value is larger, it indicates that feature items are distributed in fewer categories and have strong classification ability. If the value is small, the feature item is not representative. The specific calculation definition is shown in

$$\text{TF}_{ij} = \bar{\text{tf}}_{ij}, \tag{3}$$

$$\text{IDF}_i = \log\left(\frac{N}{n_i} \times \frac{m}{m_i}\right). \tag{4}$$

TF_{ij} represents the word frequency of feature i relative to category j. The higher the value of TF_{ij}, the better the feature i can represent this category. IDF_i represents the distribution ratio of feature i among categories. The higher the IDF_i value, the stronger the category discrimination ability of characteristic i.

(3) Word distance factor

The word distance indicates the difference between the last appearance position and the first appearance position of the feature item in the document, reflecting the range of the feature item in the text. The larger the range of feature items in this document, that is, the larger the word distance, the better it can reflect the category of the document. On the contrary, if the feature items are only distributed in a small range and the word distance is small, the category of the document cannot be well represented. If this feature item appears frequently locally in a document, the TF value will be increased, affecting the final algorithm result. Therefore, this paper introduces the word distance factor to avoid this problem. The formula for calculating the word distance factor is shown in

$$\text{WDF}_{ip} = \frac{\text{last}\left(g_{ip}\right) - \text{first}\left(g_{ip}\right) + 1}{\text{fea}(p)}, \tag{5}$$

where $\text{last}(g_{ip})$ represents the position number of the last appearance of feature item i in document p. $\text{first}(g_{ip})$ represents the position number of the first appearance of feature item i in document p. $\text{fea}(p)$ represents the total number of feature items in document p.

Considering the intraclass, interclass, and position factors of feature items, a feature item weight calculation method TF-IDF-ICP (interior factor, category factors, and position) for text classification is proposed, which is defined as

$$\text{TF} - \text{IDF} - \text{ICP}_{ip} = \text{TF}_i * \text{IDF}_i * \text{inter}\, C_i * \text{WDF}_{ip}, \tag{6}$$

where TF_i represents the word frequency of feature i relative to categories. IDF_i represents the distribution ratio of feature i among categories. $\text{inter}\, C_i$ represents the intraclass factor. WDF_{ip} represents the word distance factor. Artificial neural network can only accept numerical input, not a word as a string. In order to enable the deep learning model to process text data, first of all, it is necessary to express natural language as a numerical vector that the model can recognize. Word2vec is based on the simple shallow artificial neural network, according to the given large corpus, through training and optimization model to get the training result—word vector. This word vector can quickly and effectively express a word as a numerical vector, and can measure the similarity between words well, so as to represent the different attributes of the word.

In text classification, it is necessary to convert the phrases in the text into low-dimensional word vectors. First, the text d is segmented by jieba. Text d after segmentation is $D_i = [w_1, w_2, \cdots, w_i, \cdots, w_n]$. w_i stands for the ith word in the document. n represents the total number of words in the document. Word2vec is then used to vectorize the text. w_i is expressed as $[v_1, \cdots, v_i, \cdots, v_l]$. l is the dimension of word vector, that is, each word is represented by l dimension. Word vectors trained by Word2vec retain the relevance of words in the original corpus, but ignore the importance of different words in the text. Therefore, this paper uses the improved TF-IDF algorithm TF- IDF-ICP to calculate the weight of each word and combine it with Word2vec word vector. The specific text representation is shown in

$$\text{vec}(\text{d}) = D * \text{TF} - \text{IDF} - \text{ICP}. \tag{7}$$

D stands for Word2vec word vector.

The specific steps of weighting Word2vec are shown in Figure 1.

2.2. Dimension Reduction of Word Vector. It is a burden for the training and prediction of machine learning model to spend the conference. In this paper, the improved SPCA is designed to reduce the dimension of word vector to reduce the dimension of dataset. Then, the basic features are selected to minimize the information loss while compressing the data.

(1) Reweighted ℓ1 optimization framework

ℓ1 optimization problem originates from ℓ0 optimization problem. ℓ0 optimization problem can be expressed as follows: given an $m \times n$ matrix $\boldsymbol{A}$ and a nonzero vector $\boldsymbol{b}$, where $m \leq n$. The sparse solution of $\boldsymbol{Ay} = \boldsymbol{b}$ is solved. The mathematical form is as follows:

$$\begin{aligned} &\min_{\mathbf{y} \in \mathbf{R}^{\mathbf{n}}} \|\mathbf{y}\|_0, \\ &\text{s.t.} \mathbf{Ay} = \mathbf{b}, \end{aligned} \tag{8}$$

where $\|\boldsymbol{y}\|_0$ represents the number of nonzero solutions of $\boldsymbol{y}$. Since Equation (8) is a nonconvex optimization and NP difficult optimization problem, there is no effective solution at present, so only its approximate solution can be considered. Jojic et al. proved that the convex hull of $\|\boldsymbol{y}\|_0$ is completely ℓ1 norm. Through convex analysis, it is reasonable to use $\|\boldsymbol{y}\|_1$ to replace $\|\boldsymbol{y}\|_0$ for optimization operation, thus leading to ℓ1 optimization problem. The mathematical model for the optimization problem is shown as follows:

$$\begin{aligned} &\min_{\mathbf{y} \in \mathbf{R}^{\mathbf{n}}} \|\mathbf{y}\|_1, \\ &\text{s.t.} \mathbf{Ay} = \mathbf{b}. \end{aligned} \tag{9}$$

$\|\boldsymbol{y}\|_1 = \sum_{i=1}^{n} |\boldsymbol{y}_i|$, that is, the sum of absolute values of each element in the solution vector. ℓ1 optimization problem is a convex optimization problem that can be efficiently solved by using a convex programming tool. Through numerous experiments, it is shown for ℓ1 that by reasonably weighting their ℓ1 norm and iteratively updating the weights, the performance of their ℓ1 optimization framework is greatly enhanced. In the case of their ℓ1 norm, the larger one is punished more than the ℓ0 norm, according to the definitions for ℓ1 and ℓ0, while their ℓ0 norm treats both equally. Thus, if they add a weight matrix to the ℓ1 optimization problem, their coefficients of different sizes are punished equally. Then, their ℓ1 optimization problem will be infinitely close to their ℓ0 optimization problem, resulting in a scarcer solution. This is the advantage for a weighted ℓ1 optimization framework. The optimization problem for a weighted ℓ1 can be expressed as follows:

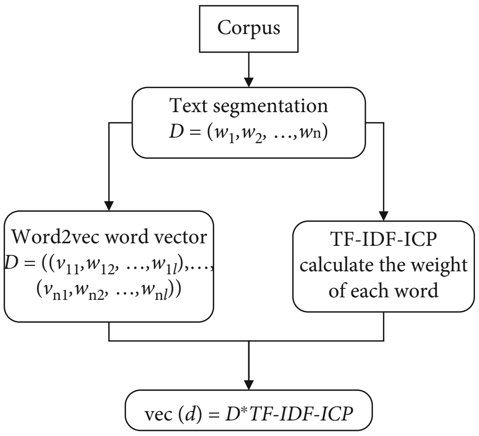
Corpus
Text segmentation
$D = (w_1, w_2, \ldots, w\text{n})$
Word2vec word vector
$D = ((v_{11}, w_{12}, \ldots, w_{1l}), \ldots, (v_{\text{n}1}, w_{\text{n}2}, \ldots, w_{\text{n}l}))$
TF-IDF-ICP
calculate the weight
of each word
vec (d) = D*TF-IDF-ICP

Figure 1: Flow chart of weighted Word2vec.

$$\min_{\mathbf{y} \in \mathbf{R}^{\mathbf{n}}} \|\mathbf{Wy}\|_1. \tag{10}$$

By selecting an appropriate weighted matrix $\boldsymbol{W}$, the reweighted ℓ1 optimization problem obtains a scarper solution for ℓ1 than the traditional ℓ1 optimization problem, thus making the result more like the ℓ0 optimization problem. The problem then turns to how to select a weighted matrix $\boldsymbol{W}$ so that the optimization problem for a reweighted ℓ1 can obtain a rarer solution while ensuring that the result is correct. According to the correlation proof, the absolute value of the weight should be inversely proportional to the value of the corresponding element in the final solution. However, if the specific value of the final solution is not known, the appropriate weighting matrix cannot be selected. At the same time, if the appropriate weighting matrix cannot be selected, the correct final solution cannot be obtained. A reasonable choice to solve this problem is to use an iterative approach, that is, by setting the weighted matrix as the identity matrix at the beginning, they can obtain an approximate solution of the reweighted ℓ 1 optimization problem. In the second iteration, the weighted matrix can be updated according to the approximate solution obtained, and then, the process is repeated until the termination condition is satisfied.

(2) Weighted sparse principal component analysis algorithm

Compared with the traditional ℓ1 optimization framework, the reweighted ℓ1 optimization framework can obtain sparser solutions. Therefore, in this paper, a new mathematical model is presented by introducing the reweighted ℓ1 optimization framework and LASSO regression model into the principal component analysis algorithm. This is the reweighted sparse principal component analysis algorithm which aims to obtain more sparse solutions.

Weighted sparse principal component analysis algorithm is an optimization model based on principal component analysis algorithm, upon which an optimization framework and LASSO regression model are added to generate sparse solutions. The mathematical model of the principal component analysis algorithm is shown in formula:

$$\underset{\mathbf{V}\in\mathbf{R}^{\mathbf{n}\times\mathbf{g}}}{\arg\min}\mathbf{Y}-\mathbf{YVV}^{\mathrm{T}}{}_{F}^{2}, \tag{11}$$

where the matrix $\boldsymbol{Y}$ is an $n \times c$ order matrix formed by the original data. The matrix $\boldsymbol{V}$ is a set of orthogonal bases. $\boldsymbol{V}^T$ represents the transpose matrix of the matrix $\boldsymbol{V}$. The principal component analysis algorithm reduces the dimension of the original data by using the base transformation method. The weighted sparse principal component analysis algorithm proposed in this paper adds an optimization frame and LASSO regression model to the principal component analysis algorithm. The mathematical model is as follows:

$$\begin{gathered}\underset{\boldsymbol{G},\boldsymbol{H}}{\arg\min}\sum_{i=1}^{n}\boldsymbol{Y}-\boldsymbol{YHG}^{T}{}_{F}^{2}+\lambda\|\boldsymbol{H}\|_{F}^{2}+\|\boldsymbol{W}\cdot\boldsymbol{H}\|_{1},\\ \text{s.t.}\boldsymbol{G}^{T}\boldsymbol{G}=\boldsymbol{I},\boldsymbol{H}^{T}\boldsymbol{H}=\boldsymbol{I},\end{gathered} \tag{12}$$

in which $\boldsymbol{Y}$ is an $n \times c$ matrix. $\boldsymbol{G}$ and $\boldsymbol{H}$ are orthogonal matrices of order $c \times d$. The matrix $\boldsymbol{G}^T$ represents the transposition of the matrix $\boldsymbol{G}$ in which $\boldsymbol{G}_{c\times d}=[\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_d]$. $\boldsymbol{H}_{c\times d}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_d]$. $\boldsymbol{W}$ is a weighted matrix of order $c\times c$, and the matrix $\boldsymbol{W}$ is a diagonal matrix. λ is the regularization coefficient. After expanding the norm in formula (12), formula (13) can be get:

$$\underset{\mathbf{G},\mathbf{H}}{\arg\min}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{y}_{\mathbf{i}}-\mathbf{HG}^{\mathbf{T}}\mathbf{y}_{\mathbf{i}}{}^{2}+\lambda\sum_{j=1}^{d}\mathbf{h}_{\mathbf{i}}{}^{2}+\sum_{j=1}^{d}\mathbf{W}\cdot\mathbf{h}_{\mathbf{i}1}. \tag{13}$$

$\boldsymbol{W}$ is the weighted matrix, and $\boldsymbol{h}_i$ is the column vector constituting matrix $\boldsymbol{H}$. According to the correlation between eigenvalues and eigenvectors, we can get:

$$\mathbf{We}_{\mathbf{i}}=\varepsilon\mathbf{e}_{\mathbf{i},} \tag{14}$$

where ε is the eigenvalues corresponding to the vector $\boldsymbol{e}_i$. According to formula (14), formula (13) can be expanded as follows:

$$\begin{gathered}\underset{\boldsymbol{G},\boldsymbol{H}}{\arg\min}\sum_{i=1}^{n}\boldsymbol{y}_{\mathbf{i}}-\boldsymbol{HG}^{T}\boldsymbol{y}_{\mathbf{i}}{}^{2}+\lambda\sum_{j=1}^{d}\boldsymbol{h}_{\mathbf{i}}{}^{2}+\sum_{j=1}^{d}\varepsilon_{j}\boldsymbol{h}_{i1},\\ \text{s.t.}\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{G}=\boldsymbol{I},\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}=\boldsymbol{I}.\end{gathered} \tag{15}$$

When solving the mathematical model shown in Equation (15), it can be solved by alternating minimization method. The idea of alternating minimization method is to first assume that two initial matrices are given and use these two initial matrices for iterative calculation. During iteration, the matrix generated by the previous iteration is used to solve the matrix of the current iteration. That is, starting from any $(\boldsymbol{G}_0,\boldsymbol{H}_0)$, matrices $\boldsymbol{G},\boldsymbol{H}$ are the obtained matrices. The matrix subscript represents the number of iterations. First, an initial matrix $\boldsymbol{G}_0$ or $\boldsymbol{H}_0$ is given, and then, iterative calculation is carried out. At the pth iteration, assuming a given matrix $\boldsymbol{G}_{(p-1)}$, the matrix $\boldsymbol{G}_{(p-1)}$ is used to solve the matrix $\boldsymbol{H}_{(p)}$. Then, the matrix $\boldsymbol{H}_{(p)}$ is used to solve the matrix $\boldsymbol{G}_{(p)}$. Then, repeat the process until the number of iterations meets the preset stop conditions.

First, consider the case that the orthogonal matrix $\boldsymbol{G}$ is known. When the orthogonal matrix $\boldsymbol{G}$ is known, only the matrix $\boldsymbol{H}$ needs to be solved in Equation (15). Therefore, the problem of solving the mathematical model shown in Equation (15) can be transformed into the problem of solving the mathematical model shown in

$$\begin{gathered}\underset{\mathbf{H}}{\arg\min}\mathrm{Y}-\mathbf{YHG}^{\mathbf{T}}{}_{F}^{2}+\lambda\|\mathbf{H}\|_{F}^{2}+\|\mathbf{W}\cdot\mathbf{H}\|_{1},\\ \text{s.t.}\mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I},\mathbf{H}^{\mathbf{T}}\mathbf{H}=\mathbf{I}.\end{gathered} \tag{16}$$

To solve the mathematical model shown in Equation (16), a new matrix $\boldsymbol{G}_{\perp}$ needs to be introduced. The matrix $\boldsymbol{G}_{\perp}$ is an orthogonal matrix. Therefore, matrix $[\boldsymbol{GG}_{\perp}]$ is an orthogonal matrix of order $c \times c$. By projecting the rows of $\boldsymbol{Y}-\boldsymbol{YHG}^{T}$ onto matrix $\boldsymbol{G}$ and matrix $\boldsymbol{G}_{\perp}$, formula (17) can be get:

$$\begin{gathered}\mathbf{Y}-\mathbf{YHG}^{\mathrm{T}}{}_{F}^{2}=\left(\mathbf{Y}-\mathbf{YHG}^{\mathrm{T}}\right)\mathbf{G}_{\perp}{}_{F}^{2}+\left(\mathbf{Y}-\mathbf{YHG}^{\mathrm{T}}\right)\mathbf{G}_{F}^{2}=\mathbf{YG}_{\perp}{}_{F}^{2}\\ +\|\mathrm{Y}\mathbf{G}-\mathbf{YH}\|_{F}^{2}.\end{gathered} \tag{17}$$

Because $\boldsymbol{YG}_{\perp}{}_{F}^{2}$ is independent of $\boldsymbol{H}$, it is not necessary to consider $\boldsymbol{YG}_{\perp}{}_{F}^{2}$ when solving the matrix $\boldsymbol{H}$. Therefore, formula (15) can be deduced as shown

$$\underset{\mathbf{H}}{\arg\min}\|\mathbf{YG}-\mathbf{YH}\|_{F}^{2}+\lambda\|\mathbf{H}\|_{F}^{2}+\sum_{j=1}^{d}\varepsilon_{j}\mathbf{h}_{j_1}. \tag{18}$$

Formula (18) is an elastic net regression problem, which can be solved by using LARS-EN (least angle regression-elastic net) algorithm [16]. At this point, when the matrix $\boldsymbol{G}$ is given, the matrix $\boldsymbol{H}$ can already be solved. At this time, the problem is how to solve the matrix $\boldsymbol{G}$ when the matrix $\boldsymbol{H}$ is given.

When the orthogonal matrix $\boldsymbol{H}$ is given, the orthogonal matrix $\boldsymbol{G}$ needs to be solved. Then, the mathematical model shown in formula (15) can be transformed into

$$\underset{\mathbf{G}}{\arg\min}\mathbf{Y}-\mathbf{YHG}^{\mathbf{T}}{}_{F}^{2}\ \text{s.t.}\mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I}. \tag{19}$$

Formula (19) is the mathematical model of the principal component analysis algorithm, so it can be solved by the solution based on the basis transformation and covariance matrix or by the method based on singular value decomposition. In this paper, the method based on singular value decomposition is used to compute the singular value decomposition of $\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{YH}$. So, $\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{YH}=\boldsymbol{U\Sigma V}^{\boldsymbol{T}}$ gives us the orthogonal matrix $\boldsymbol{G}=\boldsymbol{UV}^{\boldsymbol{T}}$. The final solution can be obtained by repeated iteration using the above method. The solution steps of the model are as follows:

Step 1. Firstly, according to the principal component analysis algorithm model shown in formula (18), the first d principal component vectors $[\alpha_1,\alpha_2,\cdots,\alpha_k]$ of matrix $\boldsymbol{Y}$ are calculated by singular value decomposition method. $\boldsymbol{Y}$ is the matrix of raw data. d is a constant value set when solving

Step 2. Initializes the matrix $\boldsymbol{G}$ to the principal component vector calculated in step 1, where $\boldsymbol{G} = [\alpha_1, \alpha_2, \cdots, \alpha_k]$

Step 3. According to the given matrix $\boldsymbol{G}$, use LARS-EN algorithm to solve formula (18), and get the matrix $\boldsymbol{G} = [\beta_1, \beta_2, \cdots, \beta_k]$

Step 4. According to the calculated matrix $\boldsymbol{H} = [\alpha_1, \alpha_2, \cdots, \alpha_k]$, update the matrix $\boldsymbol{G}$ by singular value decomposition according to formula (19)

Step 5. Repeat steps 3 and 4 until the termination conditions are met, and the final result will be obtained

Word2vec model training sets the dimension of 400, and the obtained word vector is also 400. However, maintaining the congress brings a burden to the training and prediction of deep learning model. In this paper, the reweighted SPCA algorithm is used to reduce the dimension of word vector to reduce the dimension of dataset. Then, the basic features are selected to minimize the information loss while compressing the data.

The relationship between word vector dimension and variance value of principal components is shown in Figure 2, which indicates that the first 100 dimensions can already contain most of the information of the original data, so the first 100 dimensions of word vector data are selected as the input of the model for training.

2.3. Deep Learning Model. Text is transformed into distributed word vector based on word2vec. Then, the dimension reduced word vector is obtained by reweighting SPCA algorithm as the input data of deep learning model. Text data belongs to time series data, and the emergence of each word depends on its previous word and the latter word. Recurrent neural networks are generally selected for training because of this dependence.

The structure of recurrent neural network (RNN) is different from that of general neural network, which usually consists of input layer, hidden layer, and output layer. There may be multiple hidden layers. On the other hand, the cyclic neural network adds a cyclic structure on the basis of the three-layer structure (see Figure 3). The left side of the equal sign is folded. The right side of the equal sign is an expansion. Xt is the input layer. Ht is the output layer. A is the hidden layer. Each A can be regarded as a neuron, and each neuron stores the previously input state first. After the operation, some relations between the current input and the previous input are retained, thus having the function of "memory." In this way, the previously calculated information can be captured, and the influence of the previously input data on the later data can be retained. Good timing is seized.

Long-short-term memory network (LSTM) evolved from RNN. LSTM avoids the problem of long-term dependence by deliberately designing and calculating the hidden layer state. Both RNN and LSTM have a chain structure of repeating artificial neural network modules. However, there are four neural network layers that interact with each other in a special way within each repeating module (see Figure 4).

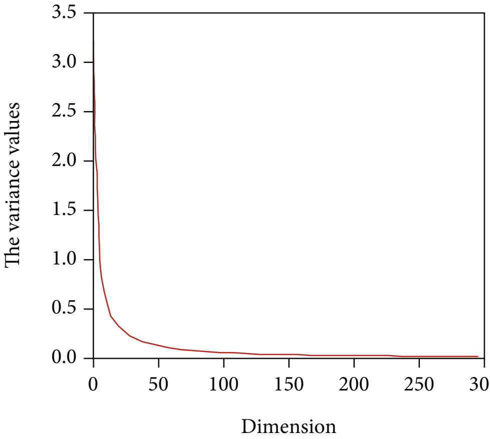

Figure 2: Relationship between dimension and reweighted SPCA variance.

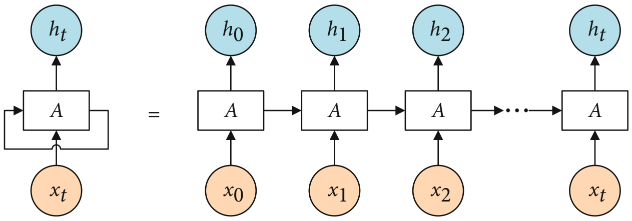

Figure 3: Structure of recurrent neural network.

In LSTM, "memory" is called cell state. The state of cells runs like a conveyor belt on the chain structure, and there is only a small amount of linear operation, which makes the information fidelity when it flows through the chain structure. In addition, LSTM adds information through a structure called a gate, which consists of a single sigmoid neural network layer and a single point multiplication operation, to capture long-term dependencies. The value range is 0~1, and whether the component information passes or not is controlled according to the numerical value. LSTM has three gates for protecting and controlling the cell state. The first gate is the "Forgetting Gate," which is implemented by the sigmoid layer and selectively forgets the information in the cell state. The second gate is the "input gate," and the sigmoid layer of the input gate determines which values to update. The candidate vector Ct created by the subsequent tanh layer selectively records the new information into the cell state. The third gate is the "output gate," whose object is the hidden layer ht, and the output part of the hidden layer is determined through the sigmoid layer. Then, it passes through the tanh layer, gets a value in (-1, 1), and multiplies it with the output value of sigmoid layer to determine the information to be output. In the LSTM network, the information in memory is selected to be retained or deleted through three gates. And the previous state, current memory, and input are combined. This structure has proved to be very effective in capturing long-term dependence. Compared with a single LSTM model, Bi-LSTM model utilizes the forward correlation information between the data

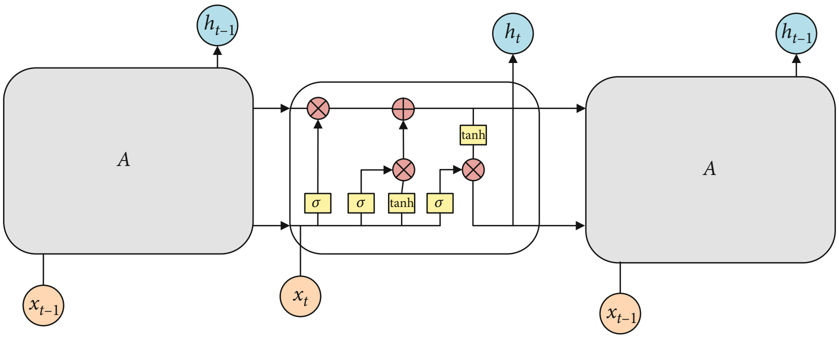

Figure 4: Repetitive module structures in LSTM.

at the time before and after the time series. Also, the reverse correlation information of the time before and after is considered. Therefore, it shows superior performance in the classification of time series. In this paper, Bi-LSTM model is selected as the training model.

2.4. The Emotional Calculation Method of Improve SPCA-LSTM. The overall framework of the improved SPCA-LSTM emotion calculation method proposed in this paper is shown in Figure 5.

(1) Input layer: emotional text data of rural tourists

(2) Text representation layer: it is to embed the words formed by combining Word2vev with TF-IDF-ICP algorithm for vector representation

(3) Dimension reduction layer: the word vector is reduced in dimension by using the reweighted SPCA algorithm, which is used to reduce the dimension of the dataset. Dimension reduction vector is used as the input of depth model

(4) Bi-LSTM layer: Bi-LSTM extracts the context semantics of text. Unidirectional LSTM generally captures the past state information and calculates the output of the current time. However, in many problems, the output of the current moment is related not only to past information but also to future information. Therefore, Bi-LSTM structure needs to be used. Bi-LSTM consists of two reverse lstms. The output at each time is determined by the forward output and the reverse output

(5) Attention layer: the attention network calculates the attention weight of each word and focuses the output features of Bi-LSTM network on the information that is more important for the current task

(6) CNN layer: put the output feature vectors into convolutional neural network to further extract local features through convolution and pooling operations, in which the convolution kernel is set to several different sizes to extract local features with different granularity

(7) Output layer: text feature vectors obtained by convolutional neural network layer are classified by softmax function

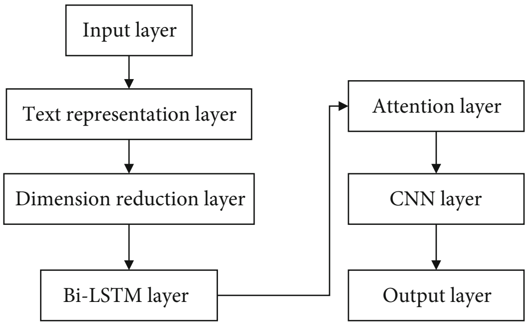

Figure 5: The emotional calculation method of improve SPCA-LSTM.

3. Experimental Results and Analysis

3.1. Experimental Environment and Dataset. The experimental environment of this paper is Tensorflow2.0 and python3.6. The compilation platform is 64-bit Spider. The hardware environment uses 8 core CPU, 8 G of memory, and NVIDIA (R)GTX (R)1080TI graphics card. Toolkits used include jieba, Scikitlearn, numpy, and gensim. The dataset is the rural tourism comment data captured from the Internet as the dataset of this paper. Among them, there are 6322 positive emotion samples and 1444 negative emotion samples, totaling 7766 samples. 80% of the data was used as training set and 20% as test set. The training set was iterated for 20 times, and the accuracy and loss function values were predicted during the training.

3.2. Experimental Parameters. The settings of model parameters in this paper are listed in Table 1 and Table 2.

3.3. Experimental Results and Comparison. Firstly, in order to verify the performance of emotion calculation method based on deep neural network proposed in this paper, the text method is compared with emotion calculation method based on emotion dictionary and emotion calculation method based on machine learning. Specific comparison methods are based on emotion dictionary ER (emotion-rules), machine learning SVM (support vector machines), and machine learning NBC (naive Bayesian classifier). The test results of the emotional model are shown in Table 3.

Table 1: Setting of word vector training parameters.

Superparameter	Parameter description	Settings
Embedding_size	Word vector dimension	100
Seq Length	Fixed sentence length	600
Num_classes	Number of label categories	10
Vocab_size	Vocabulary	8000
Window	Maximum distance between the current word and the predicted word	5
Min_count	Word frequency	1
Workers	Parallelism number of training	6

Table 2: Parameters of deep learning model.

Superparameter	Parameter description	Settings
Num_filters	Number of convolution kernels	128
Filter_sizes	Different convolution kernels	2, 3, 4
Atten_size	Attention layer size	50
Keep_prob	The probability that the neuron is retained	0.5
Learning rate	Learning rate	$1 * 10^{-3}$
Lr_decay	Learning decay rate	0.9
L2 reg lambda	Regularization coefficient	0.01
Batch_size	Training size per batch	64
Hidden_dim	Hidden layer dimension	128
Activation function	Nonlinear function	ReLU

From Table 3, the tourist sentiment calculation method based on machine learning is better than that based on sentiment dictionary. The reason is that machine learning uses statistical methods to extract feature items from text data, and its nonlinear characteristics improve the reliability and accuracy of emotion calculation. In the machine learning method, NBC is better than SVM in all evaluation indexes. And the effect on training set and test set is relatively stable. The reason is that NBC, as a classic classification model in machine learning, has a solid mathematical foundation and stable classification efficiency.

Compared with ER, SVM, and NBC, the accuracy, recall, and $F1$ values of the improved SPCA-LSTM algorithm in this paper are greatly improved. The main reason is that Bi-LSTM model with attention mechanism is intelligent in text data feature extraction and learning methods. Deep learning relies on big data and many parameters to automatically fit nonlinear prediction functions, emphasizing the depth of model structure and highlighting the importance of feature learning. At the same time, the feature representation of samples in the original space is transformed into a new feature space by feature transformation layer by layer. Compared with the method of artificial feature construction by machine learning, the algorithm model in this paper fully embodies its advantages of big data.

In addition, word2vec uses high-dimensional vectors to convert words into real vectors and accurately retain their semantic information. Then, reweighted SPCA is used to reduce the vector dimension and reduce the computation of the deep learning model. The deep learning model not only gives full play to the strong processing ability of deep learning for high-dimensional data but also preserves the good timing of tourism text data. Therefore, the proposed model achieves better results than those based on emotion dictionary and machine learning.

Table 3: Test results of emotion calculation models with different methods.

Evaluation model	Accuracy	Precision	Recall	$F1$
Emotion dictionary ER	81.0	87.1	73.4	79.6
Machine learning SVM	73.0	76.5	91.6	83.3
Machine learning NBC	83.6	90.2	90.8	90.4
Proposed model	97.4	97.8	97.2	97.5

In order to verify the effectiveness of this algorithm, this paper selects several deep learning models for comparison.

(1) CNN: TextCNN model uses Word2vec model to train word vector as word embedding layer

(2) LSTM: a network text classification model based on one-way long-short-term memory

(3) RNN+ Attention: a text classification model based on long-short-term memory based on attention mechanism

(4) BBGA: Bert model is used to train word vectors as word embedding layer, and Bi-GRU and attention mechanism are used to extract features

(5) MLCNN: a text classification model combining convolutional neural network and long-short-term memory network

(6) CTMWT: a convolutional neural network text classification model based on Word2vec and improved TF-IDF

(7) BGRU-CNN: text classification model based on recurrent neural network variant and convolutional neural network

The experimental results of each model are shown in Table 4.

The accuracy of this model is 5.44% higher than that of the traditional convolutional neural network. The precision rate is 5.0% higher. The recall rate is 5.49% higher. The $F1$ value is 5.19% higher. The highest accuracy rate of other text classification models is 96.7%. The highest precision rate is 96.5%. The highest recall rate is 96.6%. The highest $F1$ value is 96.6%. The accuracy of this classification model is 0.65% higher. The precision rate is 1.4% higher. The recall rate is 0.59% higher. The $F1$ value is 0.89% higher. Experimental data show the superiority of this method.

Table 4: Comparison of experimental results of each model.

Evaluation model	Accuracy	Precision	Recall	$F1$
CNN	92.1	92.9	91.7	92.3
LSTM	90.1	88.0	90.2	89.0
RNN + attention	93.4	93.2	93.3	93.3
BBGA	94.9	94.7	94.7	94.7
MLCNN	96.7	96.5	96.6	96.6
CTMWT	96.2	96.1	96.4	96.2
BGRU-CNN	95.2	94.3	94.5	94.4
ABLCNN+TFIDF	95.2	94.1	94.9	94.5
Proposed model	97.4	97.9	97.2	97.5

4. Conclusion

The results of big data analysis of rural tourism are introduced into the development planning of rural tourism, which will provide extremely beneficial strategic guidance for the realization of rural revitalization strategy. In order to realize the big data analysis of rural tourism, this paper proposes a rural tourism sentiment calculation method based on the improved SPCA-LSTM algorithm, which uses the improved TF-IDF+ Word2vec model to represent rural tourism data in vector. Vector dimension is reduced by reweighted SPCA. Then, Bi-LSTM model with attention mechanism is used to extract text features. Finally, SoftMax function is put into practice to calculate the emotion of rural tourists. The experiment demonstrates that the proposed algorithm is feasible and effective for sentiment analysis of rural tourists. In the next step, this paper can consider the construction of tourism-specific emotion dictionary and combine tourism-specific emotion dictionary with machine learning and deep learning methods to study tourist emotion computing.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no competing interests.

Acknowledgments

This work was supported in part by the Heilongjiang Art and Science Planning Project "Research on the Integrated Development of Music Culture and Tourism Industry in Heilongjiang Province" (No. 2021B009).

References

[1] J. Yang, R. Yang, M. H. Chen, C. H. (. J.). Su, Y. Zhi, and J. Xi, "Effects of rural revitalization on rural tourism," *Journal of Hospitality and Tourism Management*, vol. 47, pp. 35–45, 2021.

[2] Y. Liu, Y. Zang, and Y. Yang, "China's rural revitalization and development: theory, technology, and management," *Journal of Geographical Sciences*, vol. 30, no. 12, pp. 1923–1942, 2020.

[3] W. T. Fang, *Rural Tourism[M]//Tourism in Emerging Economies*, Springer, Singapore, 2020.

[4] J. M. G. Martínez, J. M. M. Martín, J. A. S. Fernández, and H. Mogorrón-Guerrero, "An analysis of the stability of rural tourism as a desired condition for sustainable tourism," *Journal of Business Research*, vol. 100, pp. 165–174, 2019.

[5] H. Zhu and F. Deng, "How to influence rural tourism intention by risk knowledge during COVID-19 containment in China: mediating role of risk perception and attitude," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, p. 3514, 2020.

[6] C. Martínez-Hernández, C. Mínguez, and C. Yubero, "Archaeological sites as peripheral destinations. Exploring big data on fieldtrips for an upcoming response to the tourism crisis after the pandemic," *Heritage*, vol. 4, no. 4, pp. 3098–3112, 2021.

[7] X. Zhang, L. Yu, M. Wang, and W. Gao, "Fm-based: algorithm research on rural tourism recommendation combining seasonal and distribution features," *Pattern Recognition Letters*, vol. 150, pp. 297–305, 2021.

[8] M. A. Tian and X. I. E. Yanjun, "The study of emotion in tourist experience: current research progress," *Tourism and Hospitality Prospects*, vol. 3, no. 2, pp. 82–101, 2019.

[9] J. C. Cuizon and C. G. Agravante, "Sentiment analysis for review rating prediction in a travel journal," in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pp. 70–74, Seoul, Republic of Korea: ACM, 2020.

[10] G. Gupta and P. Gupta, "Twitter mining for sentiment analysis in tourism industry," in *2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*, pp. 302–306, London, UK, 2019.

[11] C. Song, X. K. Wang, P. Cheng, J. Q. Wang, and L. Li, "SACPC: a framework based on probabilistic linguistic terms for short text sentiment analysis," *Knowledge-Based Systems*, vol. 194, p. 105572, 2020.

[12] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access*, vol. 7, pp. 43749–43762, 2019.

[13] L. Chen and S. Tang, "Physical-layer security on mobile edge computing for emerging cyber physical systems," *Computer Communications*, vol. PP, no. 99, pp. 1–12, 2022.

[14] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020.

[15] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image-text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[16] J. Lu and M. Tang, "Analytical offloading design for mobile edge computing based smart internet of vehicle," *EURASIP Journal on Advances in Signal Processing*, vol. PP, 10 pages, 2022.