Exploring destination image through online reviews: an augmented mining model using latent Dirichlet allocation combined with probabilistic hesitant fuzzy algorithm

Received 15 July 2021
Revised 12 September 2021
Accepted 12 October 2021

Yuyan Luo, Tao Tong, Xiaoxu Zhang, Zheng Yang and Ling Li
College of Management Science, Chengdu University of Technology, Chengdu, China

Abstract

Purpose – In the era of information overload, the density of tourism information and the increasingly sophisticated information needs of consumers have created information confusion for tourists and scenic-area managers. The study aims to help scenic-area managers determine the strengths and weaknesses in the development process of scenic areas and to solve the practical problem of tourists' difficulty in quickly and accurately obtaining the destination image of a scenic area and finding a scenic area that meets their needs.

Design/methodology/approach – The study uses a variety of machine learning methods, namely, the latent Dirichlet allocation (LDA) theme extraction model, term frequency-inverse document frequency (TF-IDF) weighting method and sentiment analysis. This work also incorporates probabilistic hesitant fuzzy algorithm (PHFA) in multi-attribute decision-making to form an enhanced tourism destination image mining and analysis model based on visitor expression information. The model is intended to help managers and visitors identify the strengths and weaknesses in the development of scenic areas. Jiuzhaigou is used as an example for empirical analysis.

Findings – In the study, a complete model for the mining analysis of tourism destination image was constructed, and 24,222 online reviews on Jiuzhaigou, China were analyzed in text. The results revealed a total of 10 attributes and 100 attribute elements. From the identified attributes, three negative attributes were identified, namely, crowdedness, tourism cost and accommodation environment. The study provides suggestions for tourists to select attractions and offers recommendations and improvement measures for Jiuzhaigou in terms of crowd control and post-disaster reconstruction.

Originality/value – Previous research in this area has used small sample data for qualitative analysis. Thus, the current study fills this gap in the literature by proposing a machine learning method that incorporates PHFA through the combination of the ideas of management and multi-attribute decision theory. In addition, the study considers visitors' emotions and thematic preferences from the perspective of their expressed information, based on which the tourism destination image is analyzed. Optimization strategies are provided to help managers of scenic spots in their decision-making.

Keywords Online review, Destination image, Text mining, LDA model, Sentiment analysis, PHFA

Paper type Research paper

The research was funded by the Humanities and Social Sciences Program of the Ministry of Education of the People's Republic of China (Grant No. 20YJC630095), National Natural Science Foundation of China (Grant No. 71971151), China's Post-doctoral Science Fund Project (Grant No. 2018M631069), Regional Public Management Informatization Research Center Project of Sichuan Province, Key Research Bases of Philosophy and Social Sciences (Grant No. QGXH20-03), National Park Research Center Project of Sichuan Province, Social Science Key Research Base (Extension) (Grant No. GJGY2020-ZD001), Sichuan Provincial Social Science Research Planning Project (Grant No. SC21B007), Key Program of Leisure and Sports Development Research Center in Sichuan Province, Key Research Base of Humanity and Social Science, Education Department of Sichuan Province (Grant No. XXTYCY2021A01), Program of Ecological Civilization Research Center in West China, and Key College Research Base of Humanity and Social Science in Sichuan Province (Grant No. XBST2021-YB001).

Kybernetes

0368-492X
DOI 10.1108/K-07-2021-0584

1. Introduction

An appealing image of a tourist destination can attract tourists and influence their willingness to revisit. Tourist perception is an important part of the image of a tourist destination. Furthermore, tourist satisfaction is a subjective expression of tourists' emotions, which connects these individuals to the image of a tourist destination at a theoretical level. In fact, tourists and tourism destination image are closely linked (Suo *et al.*, 2019). The rapidly evolving network informatization and the vigorous development of social media provide a platform for tourists to release and disseminate all kinds of tourism-related information (Liang *et al.*, 2018; Palau-Saumell *et al.*, 2016). Tourists' online reviews have a profound impact on the communication and development of tourism destination image, which contains tourists' perceptions and emotions about scenic spots (Fang and Qu, 2018). The online reviews of tourists provide new data sources and perspectives for the study of tourism destination image (Young-Seok and Kim, 2016). Some scholars have constructed basic tourism analysis lexicons to solve the problems of omission, implied emotion and polarity transfer in review texts. However, the results of the extraction of review texts in Chinese may still deviate from the meanings expressed by the tourists themselves (Guo *et al.*, 2021). The study of tourism destination image is a challenging topic at present (Zhong *et al.*, 2015). In addition, instrumental reviews significantly influence tourists' behavior; thus, tourism destination image must be examined on the basis of tourists' online reviews (Cheng *et al.*, 2016).

The study of tourism destination image began in the 1970s and has gradually gained widespread attention. Since Hunt's doctoral thesis, "A factor in the development of image tourism," which explored the significance of image factors in the development of tourism destinations, research on the image of tourism destinations has experienced an international boom. Studies have analyzed the positive and negative perceptual factors, tourists' motivation and behavior and perceptions of different types of tourists from the perspective of online reviews. Researchers have also examined tourism destination image in terms of city image perception factors, overall image perception and representative attractions (Wan *et al.*, 2017). However, tourism destination image is not materialistic and fixed, and the existing studies have not taken into account the differences between the types of tourism destinations. Furthermore, researchers have studied urban tourism destination image in a generalized manner, which has resulted in overgeneralization and locational fallacy. At the same time, the research on destination image is still in its primary stage, and the existing research mainly focuses on the destination image based on the basic data, interest data and behavior data of tourists; moreover, the multi-dimensional and comprehensive tourist portrait model and its application are insufficient (Guo *et al.*, 2020). With the widespread use of online tourism platforms, studies on online reviews have emerged and new analytical techniques have gained importance. However, improving the practical application of related methods in the study of scenic-destination image remains a challenge. Therefore, from the perspective of tourists' online reviews, scholars must optimize online review analysis techniques and analyze the destination image of scenic spots through online reviews. They must also evaluate the service management quality of tourist destinations and determine tourists' needs. These improvements can aid in the development of the topographical image of Chinese destinations.

In the era of information overload, researchers need to explore ways to help the managers of scenic spots determine the advantages and disadvantages in the development process of these spots in the face of the intensive and increasingly mature consumer demand for information. Moreover, experts need to solve the practical problems that prevent tourists from quickly and accurately obtaining the destination image of scenic spots and finding locations that meet their demand. AAAAA (5A) is the highest level of China's tourist attractions. This level comprises China's world-class boutique tourist attractions. As one of the 5A scenic spots, Jiuzhaigou's tourism industry is developing rapidly. However, in this process, problems in service quality and safeguard measures still emerge and thus negatively influencing the development of the tourism economic industry (Xi *et al.*, 2020). Therefore, this process must be improved urgently to promote the sustainable growth of the tourism industry and provide decision-making reference for other scenic spots. As such, the main objectives of this study are as follows:

(1) This work enriches the research on tourism destination image from the perspective of tourist reviews.

(2) It uses the improved LDA model, sentiment analysis and PHFA to build a complete set of text mining model for online tourism reviews, which can solve the current problems of over-generalization and insufficient depth of tourism destination image.

(3) Information overload leads to the difficulty for scenic-spot managers to determine the advantages and disadvantages in the development process of scenic spots. In addition, tourists are unable to grasp the destination image of scenic spots quickly and accurately and find the scenic spots that meet their needs. This study provides an idea and a research paradigm to solve this.

To address the above issue, this work first constructs a complete destination image analysis process. Then, it takes Jiuzhaigou as an example to analyze the following. (1) Web crawler is used to gather tourists' online comments. (2) Then, LDA is applied to explore the highlights of comments, understand the perception dimension of visitors and strengthen the identification of negative subjects. (3) Afterward, the emotional perception of tourists is analyzed using a dictionary based on the emotional analysis method to understand the change trend of tourist satisfaction. (4) Then, integrating PHFA and combining the LDA and sentiment analysis results, the attribute and its elements are analyzed from two dimensions of time and theme. (5) Taking Jiuzhaigou as an example, this study puts forward some suggestions for the operation and management of the scenic spot to help managers further understand the views of tourists. On the one hand, this study provides decision support for the management and planning of scenic spots. On the other hand, it allows tourists to select scenic spots by providing a more accurate reference.

2. Literature review

In this study, CiteSpace, a bibliometric tool, was used to select information on the "highly cited in the field" and "hot spots in the field" from the core database of Web of Science with the themes of "portrait" and "tourism." The tool was also used to locate keywords. By the end of December 2020, a total of 821 data samples were obtained. Keywords are an important sign of research hotspots. In this study, the keyword map was selected to display the high-frequency words in the keywords. This map was manually adjusted to obtain Figure 1. Large nodes indicated that the keyword appeared numerous times. As can be seen from the figure, the research on the destination image of scenic spots is relatively extensive. It is generally related to measuring tourists' satisfaction and exploring the related attributes of tourists' perception. Next, the key nodes were further analyzed through the relevant literature.

2.1 Online tourism review methods and application

Tourism is a planned activity, which involves the decision-making and planning and arrangement of food, housing, transportation, tourism, shopping and entertainment (Sun *et al.*, 2018). Tourists, especially young and middle-aged tourists, often make a detailed travel plan before traveling to maximize and enjoy the tourism experience. Travel decisions and

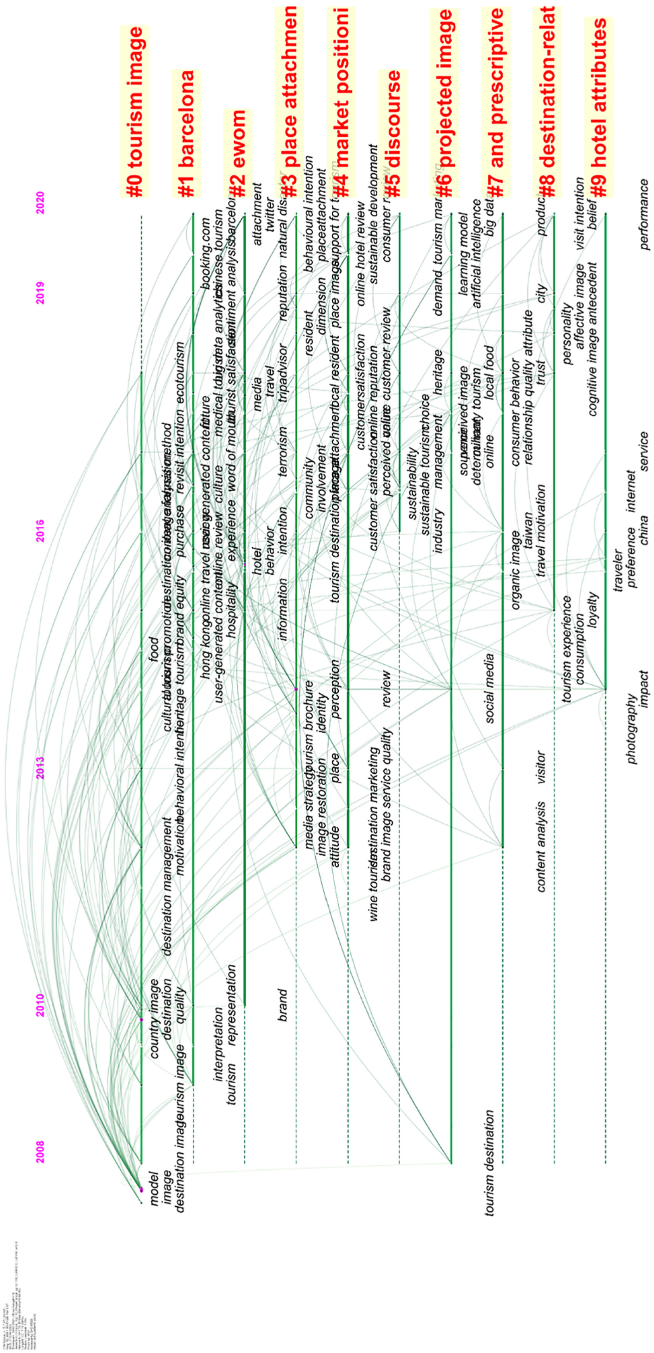

Figure 1.
Map of keyword knowledge on scenic-spot destination image research

plans need a lot of relevant information. In the early days, this information was often difficult to obtain. However, with the progress of science and technology and the popularity of the Internet, a vast majority of relevant information can be obtained now through the network. In this context, online review data, blog data and other related text format data constitute a special type of big data in tourism research – online text data (Bigne *et al.*, 2020). Online reviews present the destination image through tourists' perception; in addition, these reviews indicate that the emotional attributes and cognitive attributes in the online tourism community are interrelated (Mariani and Borghi, 2021). The existence of potential value information in online text data provides support for tourism research. Managing online reviews can promote business growth. Existing studies use such reviews to measure tourist satisfaction, explore related attributes of tourist perception, evaluate and improve hotel e-word-of-mouth (Kuo *et al.*, 2019), measure tourism satisfaction (Pestana *et al.*, 2020) and improve scenic-spot management (Nadeem *et al.*, 2018).

To extract and utilize the valuable information hidden in the online text data, a variety of text mining technologies have been widely used in the research of tourism, including two typical stages: data collection and data mining (Tunali *et al.*, 2016). The first step in analyzing online reviews is to collect online text data, including travel-related reviews and blogs, from relevant social media sites. A widely used method is web crawler. It is defined as a program or software that can traverse the web and download web documents in a systematic and automatic manner (Li *et al.*, 2008). Data mining is another key stage of mining valuable information in the text after data collection. The typical technologies of the current tourism research include topic extraction and sentiment analysis.

A topic-extraction model involves modeling hidden topics in the text, which can be used to mine potential topics and the semantic structure of text; this approach is widely used in various fields (Wu *et al.*, 2010). The existing topic models are LDA, latent semantic analysis (LSA) and non-negative matrix factorization (NMF) (Gruninger, 2007). The LDA model is a three-layer "document-topic-word" Bayesian model proposed by Blei *et al.* in 2003. It uses probabilistic derivation to determine the semantic structure of a dataset to obtain the topic of a text (Jelodar *et al.*, 2019). LSA is a model based on singular value decomposition and was proposed by Kontostathisa *et al.* The model is effective in achieving a summarized extraction of document information while achieving a dimensionality reduction of the document; however, this model suffers from overfitting (Shareef *et al.*, 2015). NMF is a topic-mining method that differs from previous probabilistic views, as it models the underlying components as axes. Each document corresponds to a unique point in the underlying linear space with a geometric perspective. However, NMF does not take into account *a priori* knowledge of topic probability distributions. For example, in sports-related training texts, the probability of a sports topic occurring is certainly higher than that of a philosophical topic. This understanding comes from *a priori* knowledge, which NMF fails to capture (Chen *et al.*, 2019). Hence, LDA was selected for this study because it uses the Dirichlet distribution, which simplifies the derivation process of the model and can effectively avoid the problem of overfitting in the LSA and the lack of prior probability in the NMF. Moreover, LDA has been widely used in previous studies to determine the key dimensions from online reviews and identify influential topics from these reviews (Jarchi and Boostani, 2006). Although it has become the mainstream topic extraction model, no in-depth mining has been conducted on the probability information generated by LDA. Hence, this study attempts to use the probability information generated by LDA combined with PHFA to optimize the destination image.

Sentiment analysis originated in the late 1990s (Blood and Phillips, 1995). It is an analysis method that helps decision-makers obtain emotional information by mining and analyzing the emotional content expressed in the text (Bahrainian and Dengel, 2013). Through natural language processing of the collected text, text information is mined and the emotional tendency of the text author toward the goods or articles that the author describes is analyzed.

Moreover, the method effectively analyzes the value of different items for users and provides assistance for later recommendation and decision-making. This kind of analysis can be divided into two categories: sentiment analysis based on a dictionary and sentiment analysis based on machine learning (Zhao *et al.*, 2016). Generally, sentiment analysis can be divided into text level, sentence level and word level (Fulse *et al.*, 2017). According to the level of text analysis, sentiment analysis can be divided into coarse-grained and fine-grained analysis (Dai *et al.*, 2019). Generally speaking, coarse-grained analysis includes text-level and sentence-level sentiment analysis, while fine-grained analysis involves word-level sentiment analysis. The fine-grained text sentiment analysis extracts the polarity and tendency of emotion from several aspects through the analysis of words to obtain an accurate degree of sentiment tendency (Wei *et al.*, 2020). At present, fine-grained sentiment analysis based on a sentiment dictionary is a common method for text sentiment analysis (Song *et al.*, 2017). The sentiment dictionary sets one or more sentiment scores for specific sentiment words through sentiment polarity and sentiment score processing to achieve fine-grained sentiment analysis at the word level (Wu and Liang, 2020). On the basis of the similarity calculation method of Chinese words, scholars have proposed a method to calculate the emotional weight of Chinese emotional words; they have also built a basic Chinese emotional dictionary based on the HowNet emotional words set and have classified sentence emotions (Jiang *et al.*, 2018). The accuracy of fine-grained sentiment analysis can be increased by constructing a fine-grained sentiment word bank to analyze sentiment tendency (Haselmayer and Jenny, 2017).

2.2 Tourism destination image

The concept of destination image consists of two components: a perceptual-cognitive component that captures knowledge and beliefs about a destination's attributes and an affective component that describes feelings toward a destination (Beerli and Martin, 2004). The cognitive and affective components work to influence the overall image of a particular destination of past or prospective tourists (Baloglu and McCleary, 1999). For decades, the image of a tourist destination has been an important topic in tourism literature (Pike, 2002). Destination image is often regarded as an important aspect of tourism development and destination marketing success because they affect supply and demand marketing (Chew and Jahari, 2014). Furthermore, it provides a valuable concept for tourists in the process of destination selection (Baloglu and Mangaloglu, 2001). Tourism destination image can also help tourism-marketing managers present their own market image by combining communication channels (Colombino, 2009). Given the importance of destination image to tourism, tourism scholars have been interested in measuring this image (Ladhari and Souiden, 2020).

In recent years, numerous achievements have been made in the study of destination image. The research on the perception elements and mechanism based on the destination image of tourist attractions is an important part of the research on the destination image perception of tourist attractions (Alrawadieh *et al.*, 2019). In the era of big data, the emergence of tourists' online text information brings increased possibilities to this field. At the same time, tourist destinations are experiential products, and their quality mainly depends on service quality, which is related to tourists' online text information, such as tourist evaluation and travel notes (Moon and Han, 2019). Therefore, scholars have carried out a lot of research on the measurement of scenic spots based on tourists' online texts, and the research results include using Internet big data text mining and word frequency statistics to analyze the image of tourist destinations (Xiao *et al.*, 2020). From the perspective of tourists, the destination image is explored from three aspects of cognition, emotion and overall perception (Hou, 2017). On the basis of the use of tourists' online text information, the above research excavates the emotional features and perceptual elements in the text information to depict the destination image of scenic spots. At the same time, from individual information integration to group wisdom, the direct expansion of the sample size is often used, which lacks the consideration of the similarity and difference of group tourists, as well as the incompleteness and total hesitation of information expression.

2.3 Recent research on probabilistic hesitant fuzzy sets (PHFSs)

Two main difficulties emerge in the continuous development of new ways to express information for decision-making. The first one is the ambiguity and hesitancy at the subjective level of the expert. Among them, ambiguity refers to experts often being unable to provide a specific assessment value when evaluating the objectives. These experts are only able to offer a vague and unclear range. Hesitation refers to experts being reluctant about several possible values when evaluating the objective. The second difficulty is the different levels of importance among various affiliations. To overcome these difficulties, PHFSs are created, which add corresponding probabilistic information to each affiliation that effectively expresses the different levels of importance among affiliations.

In studies related to PHFSs, probabilistic hesitant fuzzy preference relations (PHFPRs) can provide an effective way for decision-makers to express two-by-two comparison preference information for options. In the decision-making process, consistency significantly affects the outcome of the decision. Existing studies have used multiplicative preference relations to deal with consistency, such as through an automatic iterative algorithm to check and improve the geometric consistency index (Lin *et al.*, 2018). However, the traditional model suffers from the large computation of the consistency index, and no adjustment of the PHFPRs and other defects has been made. Therefore, an automatic consistency improvement model for PHFPRs is created. The model uses probability splitting to normalize the PHFSs; the consistency index, which calculates the degree of deviation between PHFPRs and its multiplicative consistency PHFPRs, can repair the inconsistent PHFPRs, thus overcoming the defects of the traditional model and achieving improvement (Lin *et al.*, 2020). In addition, methods related to PHFSs have been widely used. (1) Distributed stream processing frameworks (DSPFs) can handle the real-time data processing and analysis of Internet of Things (IoT) applications. The use of PHFA can realize the priority ranking of DSPFs. Lin *et al.* (2021) established a DSPF hybrid evaluation criteria system consisting of qualitative and quantitative criteria. This system introduced PHFSs and proposed a new probabilistic-hesitation fuzzy multi-objective optimization based on ratio analysis with the full multiplicative form (MULTIMOORA) method to achieve the ranking of DSPFs. (2) Considering the limited rational behavior of decision-makers, such as reference dependence and loss avoidance, Liang *et al.* (2018) proposed a Pythagorean uncertainty language TODIM (an acronym in Portuguese for Interactive and Multi-criteria Decision Making) method based on generalized Choquet integral to rank the urban haze pollution control schemes and select the best scheme according to the overall perceived dominance degree.

In summary, in the actual decision-making process, PHFSs, as a practical and effective tool, can comprehensively and meticulously express and deal with the uncertain preference information of decision-makers, which can effectively enhance the rationality and credibility of decision results. With the rapid development of society, people encounter increasingly complex practical decision-making problems. The need to deal with complex qualitative decision-making problems with analytical tools that are more consistent with objective reality and human way of thinking has become increasingly urgent.

This study area has performed some in-depth mining of tourist expression information from the similarities and differences of tourist comments. The information gathered provides multi-dimensional characteristics for the analysis of the destination image of scenic spots. The existing research on the analysis of tourists' online reviews, but less from the perspective of the probability of the occurrence of different membership degrees in tourists' individual expressions, reflects the preference of tourists' individual expressions and basically regards the occurrence probability of the different membership degrees of an element belonging to a certain set as the same. Therefore, on the basis of the perspective of destination image in the process of mapping tourists' information to the image, this study introduces PHFA, considers the different preferences expressed by tourists in online reviews and aggregates multi-granularity linguistic information with various individual preferences to analyze the destination image more accurately. In this way, the research can help scenic-area managers to determine the most useful information in the overloaded data. Moreover, the study provides decision support for the managers and solves practical problems, such as tourists not being able to obtain the destination image of the scenic area quickly and accurately and not finding a scenic area that meets the demand. The research also offers more accurate scenic-area information services for tourists.

3. Methodology

Analyzing the destination image of scenic spots has comprehensive benefits in protecting the ecological environment, developing the local economy and improving the quality of tourism (Lee and Xue, 2020). National 5A scenic spots are classified according to the quality of tourist attractions in China. It has five levels, with 5A being the highest level among China's tourist attractions; this category represents China's world-class boutique tourist spots (Yin *et al.*, 2015). As one of the 5A scenic spots, Jiuzhaigou has developed rapidly, which has further promoted the development of Jiuzhaigou's tourism economy. However, this process still has issues regarding the service quality and safeguard measures of Jiuzhaigou tourism, which negatively impacts the development of the tourism economic industry (Deng *et al.*, 2020). Therefore, this process must be urgently improved to promote the sustainable growth of tourism more effectively. This study intends to build an online review-analysis framework modeled from the machine learning method. On the basis of the method of the emotion dictionary, online reviews are analyzed to understand the satisfaction of tourists. At the same time, this study uses the LDA model to explore the attributes in tourist reviews and understand tourists' perception. The PHFA and LDA models are used to analyze the attributes and the elements in the attributes from the two dimensions of hesitancy and ambiguity to understand the tendency of tourists. At the same time, taking Jiuzhaigou as the

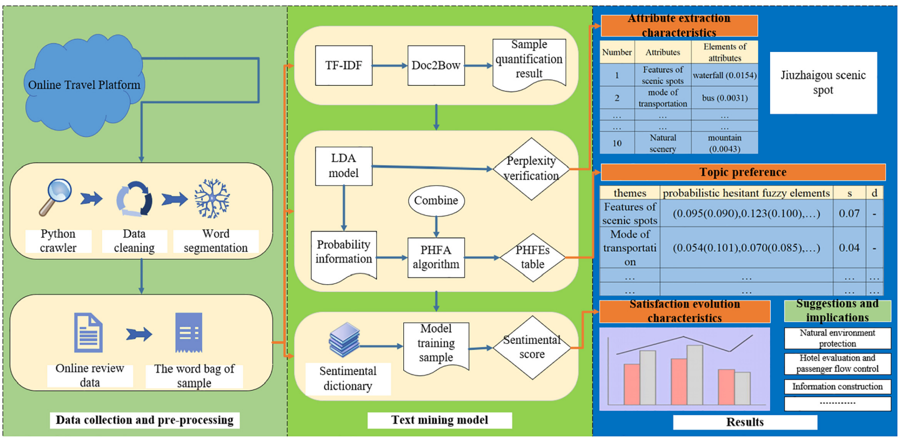

Figure 2.
Online reviews-mining framework

target scenic spot, the destination image of the scenic spot is of great reference value. This study paradigm has certain reference significance for other types of research. The overall framework of the model is shown in Figure 2.

3.1 Data collection and preprocessing

Crawler, as an old network technology, has been given increasing attention and favored by more people in the advent of the era of big data and artificial intelligence (Chan *et al.*, 2008). Python, which has sufficient functionality, stands out in web crawler technology. The data set of this study is the online comment text data of Jiuzhaigou tourism, which are unstructured and have a large sample size. The Internet crawler technology based on Python is conducive to data collection and preprocessing in this research.

The data collection and related preparations of this study are as follows. The first step is to collect data. The network comments on all dimensions and aspects of scenic spots are highly suitable for the evaluation and analysis of these destinations. The sample data used in this study are from tourism communities in China: TripAdvisor, Qunar and Ctrip Travel. As of October 2020, a total of 24,561 online reviews from 2015 to 2020 have been collected through the Python web crawler approach. The second step is to finish the preprocessing of text data. Meaningless words, inactive words (default disabled words and self-built words), punctuation symbols and other unnecessary text have been deleted from 24,222 online comments.

3.2 LDA model

Tourists select words to express their beliefs, ideas and impressions of the scenic spot. Then, they share the attributes of the destination image that they perceive through their comments on the tourism information platform. These properties cannot be observed directly. Therefore, we need to use a method to mine the attributes of the destination image outside the online comment itself.

LDA is an unsupervised machine learning technology, which can be used to identify the topic information hidden in large-scale document sets or corpora. It adopts the bag-of-words method, which regards every document as a word frequency vector, to transform the text information into digital information, which is easy to model (Yuan *et al.*, 2016). The bag-of-words method does not consider the order of words, which simplifies the complexity of the problem and provides an opportunity to improve the model. The specific structure is depicted in Figure 3.

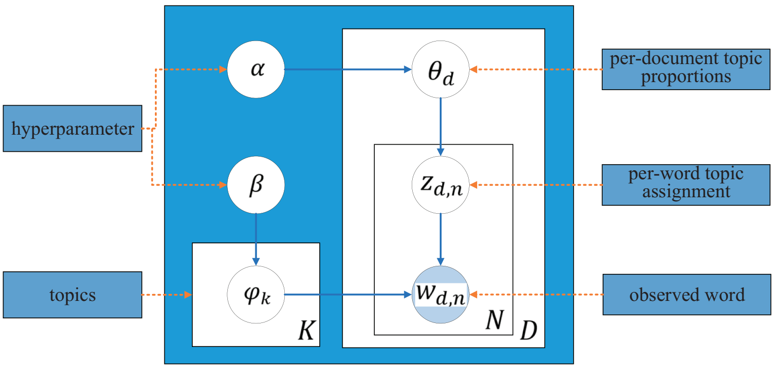

Figure 3.
LDA model

In Figure 3, the number in the lower right corner of the box is the number of repetitions of the variables contained in it. The solid circle indicates the observed values, the hollow circle indicates the implied random variables and the arrows indicate the dependencies of the variables. The referential relationships between the variables are as follows:

(1) K is the number of topics;

(2) D is the number of documents;

(3) N is the number of words in the document;

(4) $w_{d,n}$ is the nth word of the dth document;

(5) $z_{d,n}$ is the nth topic of the dth document;

(6) θ_d follows the *Dirichlet* probability distribution with hyper-parameter α;

(7) φ_k follows the *Dirichlet* probability distribution with hyper-parameter β.

In Figure 3, w is the observable variable and all others are hidden variables. Suppose a corpus is a collection of D documents, each document d is a sequence of N words, denoted as $W = [w_1, w_2, w_3, \cdots, w_n]$. The generation process of each document in corpus D is as follows:

(1) Select a topic probability distribution from the topic *Dirichilet* probability distribution $\theta \sim \mathrm{Dir}(\alpha)$;

(2) Randomly sample the words $p = (w_{d,n} \mid \theta, \varphi)$ in the document according to the corresponding word probability distribution in the obtained topic probability distribution;

(3) Repeat the above process to generate the document.

The joint probability distribution function of the LDA model is as follows:

$$p(\varphi, \theta, z, w) = \left(\prod_{i-1}^{k} p(\varphi_i \mid \beta)\right) \cdot \left(\prod_{d-1}^{D} p(\theta_d \mid \alpha) \prod_{n-1}^{N} p(z_{d,n} \mid \theta_d) \prod_{d-1}^{D} p(z_{d,n} \mid \varphi_{i:k}, z_{d,n})\right). \quad \text{(3-1)}$$

When evaluating LDA, Blei *et al.* (2002) proposed to use the degree of confusion as a criterion. The distribution of probability measures or probability models and the distribution of samples are shown below. For a text corpus containing N test documents, the number of words contained in document d is N_d and $p(d)$ is the probability of occurrence of each document in the test set. Then, the perplexity of the language model is as follows:

$$Perplexity = -\frac{1}{N} \sum_{d=1}^{N} \frac{1}{N_d} \log p(d). \quad \text{(3-2)}$$

On the basis of the LDA model, an unsupervised topic-clustering model is used to cluster the text. However, this kind of clustering requires the text to be normative and suitable for the news and other corpora processed for the first time. Online travel reviews are short, colloquial, noisy and semantically sparse. Therefore, the TF-IDF method is used to optimize this model to improve the clarity of semantic distribution (Wang *et al.*, 2016). TF-IDF is a classical statistical method to evaluate the importance of text items. This method tends to filter out high-frequency words with low resolution and retain low-frequency words with high resolution. TF is the frequency of words appearing in the document. These frequencies vary widely due to the length of different documents. They should be standardized to compare frequencies in the same environment as follows:

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3-3}$$

where tf_{ij} is the word frequency of n_i in document j, $n_{i,j}$ represents the frequency of the word n_i and $\sum_k n_{k,j}$ is the total number of words contained in document j.

IDF is the weight of a given word. Power signifies that the word is important. On the basis of word frequency, the common words are given light weight. Words with weight can be expressed as follows:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j| + 1}, \tag{3-4}$$

where $|D|$ is the total number of files in the corpus and $|\{j : t_i \in d_j\}|$ denotes the number of files containing the word t_i. Finally, the formula of TF-IDF is as follows:

$$TF - IDF = TF * IDF. \tag{3-5}$$

Therefore, in this study, an improved LDA topic-extraction model based on TF-IDF is used. Then, we modify the model parameters through the evaluation results and the perplexity output. The relationship between the output topic and the degree of confusion is shown in Figure 4. The results reveal that the effect is the best when the number of topics is ten.

3.3 Sentiment analysis

Sentiment analysis is a process of extracting affective-frame knowledge from the text and collecting affective information and data statistics on the basis of frame-semantic theory (Almars *et al.*, 2017). At present, an organic combination of text analysis and destination image technology exists and the application of the combination of the two technologies is also very extensive. Therefore, this study combines the theory of sentiment tendency analysis to extract sentiment information from online tourism reviews to provide tourists' sentiment information for the analysis of the scenic-destination image, which is conducive to obtaining accurate personalized recommendation.

The analysis methods of emotional tendency are as follows:

(1) Build a sentiment dictionary: The English text is mainly based on the expansion of English Dictionary WordNet, while the Chinese text is mainly based on the expansion of HowNet dictionary. Combining them with the characteristics of related fields, we can complete the emotion dictionary.

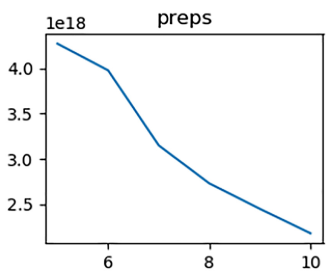

Figure 4.
Perplexity based on number of topics

(2) Emotional orientation calculation: Combined with the sentiment dictionary, the sentiment words are matched with the comment text. According to the specific algorithm, the sentiment words are given different weights, and the sentiment score of the comment text can be further obtained to complete the calculation of sentiment tendency.

(3) Determine the threshold to judge the text orientation: In this study, the text emotional tendency is divided into positive emotional tendency, negative emotional tendency and non-tendency on the basis of the current common judgment method of emotional tendency.

Sentiment analysis technology includes text pre-processing and feature word selection. Given that this research has selected the emotion analysis method based on the emotion dictionary, the feature words are extracted according to the emotion dictionary.

3.4 Integrated model based on LDA and PHFA

In the process of developing new ways to express decision-making information, two main difficulties emerge (Ren *et al.*, 2021). The first is the fuzziness and hesitation on the subjective level of experts. The second difficulty is that the importance of membership degrees varies. The emergence and development of PHFA has provided a positive way of solving the above problems.

The basic unit of PHFA is the probabilistic hesitant fuzzy element.

X is the non-empty set (Zhang *et al.*, 2017). The mapping from X to a probability distribution function on the closed interval [0,1] is the PHFS. Its mathematical expression is as follows:

$$H = \{x, h_x(p_x) \mid x \in X\} \tag{3-6}$$

where h_x is a sub-set of [0,1], which is the membership degree of $x \in X$ belonging to set E; p_x is also a sub-set of [0,1], which expresses the corresponding probability interpretation of h_x and $h_x(p_x)$ is a probabilistic hesitant fuzzy element, which is abbreviated as $h(p)$:

$$h(p) = \{h_l(p_l)|l = 1, 2, \cdots, |h(p)\ |\ \} \tag{3-7}$$

where $|h(p)|$ denotes the number of membership degrees in $h(p)$ and p_l denotes the probability corresponding to the membership degree h_l and satisfies $\sum_{l=1}^{|h(p)|} p_l \leq 1$.

PHFA can sort probabilistic hesitant fuzzy elements and eliminate uncertainty and subjectivity.

For a probability hesitant fuzzy element (PHFE) $h(p)$, its score is defined as follows:

$$s(h(p)) = \sum_{l=1}^{|h(p)|} (\gamma_l p_l) \Big/ \sum_{l=1}^{|h(p)|} p_l. \tag{3-8}$$

For two PHFEs $h_1(p)$ and $h_2(p)$, if $s(h_1(p)) > s(h_2(p))$, then we consider that $h_1(p)$ is superior to $h_2(p)$, which is denoted as $h_1(p) > h_2(p)$ or $h_2(p) < h_1(p)$. Thus, two PHFEs can be easily compared by using their scores. However, if $s(h_1(p)) = s(h_2(p))$ then, the two PHFEs cannot be ranked anymore. In this case, the deviation degrees of PHFEs are defined as follows:

$h(p) = \{\gamma_l(p_l)|l = 1, 2, \ldots, |h(p)|\}$ is a PHFE, and its score $s(h(p))$ is denoted by $\overline{\gamma}$. Then, the deviation degree of $h(p)$ is as follows:

$$d(h(p)) = \sum_{l=1}^{|h(p)|} \left(p_l(\gamma_l - \overline{\gamma})\right)^2 \Big/ \sum_{l=1}^{|h(p)|} p_l. \tag{3-9}$$

Thus, a method can be proposed to compare two PHFEs $h_1(p)$ and $h_2(p)$as follows:

(1) If $s(h_1(p)) > s(h_2(p))$, then $h_1(p) > h_2(p)$;

(2) If $s(h_1(p)) = s(h_2(p))$ and $d(h_1(p)) < d(h_2(p))$, then $h_1(p) > h_2(p)$;

If $s(h_1(p)) = s(h_2(p))$ and $d(h_1(p)) > d(h_2(p))$, then $h_1(p) < h_2(p)$;

If $s(h_1(p)) = s(h_2(p))$ and $d(h_1(p)) = d(h_2(p))$, then $h_1(p)$ is equivalent to $h_2(p)$, which is denoted as $h_1(p) \sim h_2(p)$.

In the above research, the frequency topic extracted by LDA is based on the individual expression information, which is full of uncertainty and inaccuracy. However, LDA also provides us with "subject document" probability and "subject word" probability. PHFA can use these two sets of data to aggregate individual expression information into group wisdom to express tourists' generated information better.

The specific process is as follows:

(1) LDA takes a single-comment main sentence as the calculation unit. After model training, each comment can be classified to the corresponding topic according to the topic distribution probability. In this way, we can obtain the number of comments for each topic. After normalization, we can derive the hesitation degree h of each topic (the probability of tourists discussing each topic).

(2) At the same time, the word distribution probability demonstrates the similarity between the feature words and each topic, i.e. the degree p of the comment belonging to the topic.

(3) After obtaining two sets of probability information, multiple probability hesitant fuzzy elements $h(p)$ can be constructed.

Therefore, based on the probability hesitation element, this study constructs the probability hesitation fuzzy element from the two dimensions of "time" and "theme." By calculating the probability hesitation fuzzy value, the elements in the same dimension are compared to analyze the tourists' perception of the scenic spot further. Compared with the traditional methods, the fusion of PHFA eliminates the uncertainty of the probability information generated in LDA and the inaccuracy of the individual expression information of massive reviews. Then, we can have a more accurate insight into the tourists' perception and make more correct optimization decisions and suggestions.

In summary, this study constructs an online comment-text-mining model. Compared with traditional text-mining models (mostly based on the qualitative analysis of peripatetic research data, which may lead to the omission of attributes or biased analysis results), it introduces PHFA on the basis of covering mainstream machine learning models (LDA) and mainstream sentiment analysis methods (sentiment dictionary). By extracting "topic-word" and other probabilistic data, the model can better coalesce single review texts into multiple subjects with representativeness. Then, it can analyze the image of scenic destinations based on the cognition and sentiment of multiple subjects on the attributes of scenic spots. These factors make the analysis results more objective and realistic and avoid the bias formed by individual cognitive differences. Furthermore, the model constructed in this study can provide a research paradigm for other similar studies. Tourists' online comments are brief and colloquial, with high-noise and -sparse semantic information. Therefore, on the basis of the traditional LDA method, this study uses the TF-IDF method to optimize the LDA model, which improves the clarity of the semantic distribution and enhances the accuracy of the LDA topic extraction results.

4. Results

4.1 Descriptive statistics

The obtained online tourism comment text is composed of multiple sentences and contains interference words and rare words, etc. Thus, the data needs to be preprocessed. The original 24,561 comments are processed and the effective data are 24,222 in total. Then, the effective data are divided into components by using the Jieba segmentation. The words that are used to stop and interfere in the segmentation results are screened and eliminated. Finally, the data set is loaded to obtain the final segmentation results (the statistics of high frequency segmentation results are shown in Table 1), which provides data support for subsequent LDA model application.

As shown in Table 1, the top-ten high-frequency words of Jiuzhaigou online comment include "beauty," "scenery," "worth," "scenic spot," "water," "waterfall," "good," "ticket," "suggestion" and "recommendation." To some extent, high-frequency words illustrate the tourists' attention to the scenic spot and the degree of attention. They are also closely related to the theme emotion analysis of online comments based on LDA in the later study.

4.2 Topic analysis based on LDA

In this work, the LDA model is used to obtain the last ten attributes of tourism reviews as shown in Table 2. We extract 10 attributes and 110 attribute elements. We artificially delete the meaningless attribute elements and finally retain 100 attribute elements. The first column briefly summarizes the meaning of each topic tag. The second column shows the most frequent and exclusive words related to attributes, i.e. the words that appear most often in the attributes but least frequently in other attributes. The number in brackets is the weight of the attribute, which indicates the importance of each word in the attribute. The attributes in Table 3 comprehensively reflect the attributes and elements perceived by tourists in the process of tourism. They demonstrate certain emotional attitudes, which indicates that the model proposed in this study is effective. In the following part, we will use PHFA to analyze the results further and clarify the importance of tourist tendency theme and each attribute.

Table 1. High-frequency vocabulary of Jiuzhaigou's online reviews

Number	Words	Frequency
1	beautiful	802
2	scenery	771
3	worth	662
4	scenic spot	543
5	rivers	479
6	waterfall	451
7	not bad	386
8	admission ticket	286
9	proposal	236
10	recommend	231

Number	Theme	High-frequency words (weight)
1	Features of scenic spots	Waterfall (0.015), good (0.015), Wuhuasea (0.010), Pearl (0.010), beach (0.008), recommendation (0.007), Huanglong (0.006), Panda (0.006), archery bamboo (0.005), original forest (0.004)
2	Mode of transportation	Suggestions (0.011), sightseeing (0.007), online (0.004), advance (0.004), ticket collection (0.004), bus (0.003), bus (0.003), peak season (0.003), walking (0.003), Airport (0.002)
3	Crowding degree	Off season (0.006), scenery (0.006), walking (0.006), trestle (0.005), ditch (0.005), car (0.004), bubble surface (0.003), peak season (0.003), completion (0.003), stop (0.002)
4	Weather and climate	Winter (0.008), super (0.005), delicious food (0.005), fairy tale world (0.004), two days (0.004), fish (0.004), cloud and fog (0.004), dazzle (0.004), autumn (0.003), colorful (0.003)
5	Accommodation environment	Scenery (0.025), beauty (0.022), place (0.013), beautiful (0.006), tourism (0.005), world (0.005), nature (0.004), paradise on Earth (0.004), good (0.003), mood (0.003)
6	Tour mode	Scenic spot (0.013), walk (0.009), Changhai (0.009), sightseeing bus (0.008), queue (0.007), car (0.005), plank road (0.003), management (0.003), ditch (0.003), play (0.003)
7	Tourism cost	Ticket (0.003), tour guide (0.006), peak season (0.005), play (0.005), expensive (0.004), hotel (0.004), ticket (0.004), cost performance (0.004), price (0.004), service (0.003)
8	Service experience	Jiuzhaigou (0.042), morning (0.005), tourism (0.004), Chengdu (0.004), thinking (0.004), satisfaction (0.004), autumn (0.003), feeling (0.003), regret (0.002), tourists (0.002)
9	Tour value	Worth (0.017), magic (0.007), the most beautiful (0.007), beautiful scenery (0.007), Taimei (0.006), Midea (0.006), nature (0.006), fairyland on Earth (0.004), Paradise (0.003), shock (0.002)
10	Natural scenery	Water (0.023), well-known (0.008), environment (0.006), mountain (0.004), beauty (0.004), staggered peak (0.004), Huangshan (0.003), fresh air (0.003), clear bottom (0.003), sunshine (0.003)

Table 2. Statistical table of theme distribution and high-frequency word weight

Theme \ Year	2015	2016	2017	2018	2019	2020
Features of scenic spots	990	378	450	156	372	396
Mode of transportation	558	216	216	96	210	270
Crowding degree	1,260	402	426	186	252	420
Weather and climate	2,394	390	402	330	420	558
Accommodation environment	576	132	150	96	174	228
Tour mode	522	168	180	96	162	204
Tourism cost	486	168	180	114	216	216
Service experience	954	462	276	102	192	234
Tour value	828	222	300	114	228	954
Natural scenery	1746	528	438	168	282	378
Number of comments	10,314	3,066	3,018	1,458	2,508	3,858

Table 3. Statistical table of the number of top-ten topic comments

4.3 Sentiment analysis based on dictionary

This study uses the emotional dictionary to classify emotional tendency. On the basis of the LDA theme model, we explore the emotional tendency of tourists' comments from the two dimensions of "time" and "theme." In this step, online comments are taken as a whole to judge the emotional tendency of each comment. This method reflects the tourists' comprehensive satisfaction with Jiuzhaigou as a scenic spot and involves the analysis of specific attributes.

From the "time" dimension: This study divides online reviews from 2015 to 2020 by year. The final positive comments are 20,266, the neutral comments are 148 and the negative comments are 3,808. The satisfaction is 83.68%. From Figure 5, the overall satisfaction of tourists has been fluctuating and declining this year. With the growth of population, the number of tourists in the domestic tourism market has been steadily rising. On the one hand, this increase has promoted the development of Jiuzhaigou's tourism economy. On the other hand, it also puts forward higher requirements for Jiuzhaigou. At the same time, with the development of the Internet, a growing number of people opt to express their emotions through online platforms. After a natural disaster in 2018, tourists' satisfaction with Jiuzhaigou gradually declined. The disaster caused by geomorphic characteristics can be inferred to be related to this decline. In addition, the scenic-area managers in Jiuzhaigou should actively respond to the changing needs of tourists and restore and improve the ability and quality of scenic-area operations and management.

From the perspective of "theme," as shown in Table 4, ten themes can be obtained to represent the destination image of scenic spots under the training of the LDA theme model.

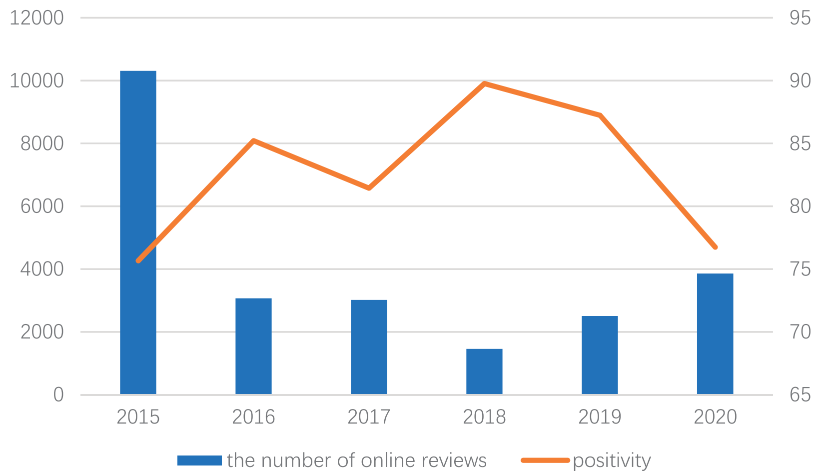

Figure 5.
Online reviews' number and trend of positivity

Table 4.
Statistical table of sentiment score of comments under subject category

Theme	Number of reviews	Average sentimental score	Number of positive polar clauses	Number of negative polar clauses
Features of scenic spots	2,742	9.8	2,251	480
Mode of transportation	1,566	1.2	1,511	52
Crowding degree	2,946	−3.4	2,044	901
Weather and climate	4,494	2.3	3,072	1,372
Accommodation environment	1,356	−1.8	1,265	85
Tour mode	1,332	0.3	1,222	102
Tourism cost	1,380	−0.9	1,322	44
Service experience	2,220	2.1	2,100	104
Tour value	2,646	5.6	2,027	599
Natural scenery	3,540	7.9	3,452	69

Among them, the three themes of “scenic spot characteristics,” “natural scenery” and “tour value” have high emotional scores, which also reflect the characteristics and competitive advantages of Jiuzhaigou’s scenic spot from the side. The emotional scores of the three themes of congestion, accommodation environment and tourism cost are negative, which indicates some problems in the management of tourism services. Combined with the tourists’ real comments, it also reveals issues on poor accommodation conditions, dense flow and congestion.

4.4 Tourist preference analysis based on PHFA

In this study, individual comments belong to the individual expression of information, with uncertainty and subjectivity. The massive individual information of LDA for text-mining topic analysis is used to obtain the comment corresponding to the “subject document” probability and “subject word” probability using PHFA. These uncertain and subjective individual data can be gathered into real and accurate group wisdom, which can then provide a more scientific and accurate expression of tourists’ wishes. The table of PHFEs is shown in Table 5. s represents the score of each PHFE while *d* represents the degree of deviation of each PHFE. Only PHFEs with the same score need to be calculated. exp (log) is the logarithmic processing of PHFEs to facilitate the comparison of PHFEs. In Table 5, two types of PHFEs are presented. The first is based on time series and each column is a temporal PHFE. The second is based on the LDA topic, where each line is a topic-fuzzy element. Combined with the score function, each PHFE is calculated through the PHFA. In the same dimension, the elements with the same score need variance calculation for further comparison.

(1) Theme orientation: On the basis of the theme dimension, the scores of the fuzzy elements of the theme probability are calculated and sorted. The results are as follows: “weather and climate” > “natural scenery” > “scenic features” > “congestion” > “tour value” > “service experience” > “transportation mode” > “tourism cost” > “tour mode” > “accommodation environment.” The clear tendency of tourists can further help scenic-area managers to build the scenic area. On the basis of the PHFA, tourists are more inclined to talk about “weather and climate,” “natural scenery” and “scenic features” under the ten themes. Combined with emotional analysis, the calculation demonstrates that Jiuzhaigou has achieved good results in three aspects of operation and management. Tourists have a high sense of identity to its natural scenery and scenic culture. However, the theme of “crowding degree” and “service experience” reflects the poor experience of the scenic spot. The managers of scenic spots should continue to maintain excellent landscape quality, strengthen the awareness of tourism services and improve the satisfaction of tourists in the process of tourism.

(2) Attitude change: On the basis of the time dimension, the time PHFE scores are calculated and sorted. The results are “2019 > 2017 > 2016 > 2018 > 2020 > 2015.” Meanwhile, on the basis of the time series analysis, the overall tendency of tourists to Jiuzhaigou fluctuates, which further verifies the fluctuation of tourists’ satisfaction with the change of time in the sentiment analysis.

4.5 Suggestions and implications

The purpose of the research on the destination image of scenic spots is as follows. (1) The study aims to achieve a better understanding of tourists and mine more information about them. Then, the research seeks to analyze the image of scenic spots from the perspective of tourists and provide recommendation strategies based on the personalized needs of tourists. (2) The research also aims to understand dynamically the problems in the operation and

Table 5.
Probabilistic hesitant fuzzy element table

Year Theme	2015	2016	2017	2018	2019	2020	s	d	exp (log)
Features of scenic spots	0.095 (0.090)	0.123 (0.100)	0.149 (0.103)	0.106 (0.103)	0.148 (0.103)	0.102 (0.101)	0.073	–	–
Mode of transportation	0.054 (0.101)	0.070 (0.085)	0.071 (0.102)	0.065 (0.103)	0.083 (0.097)	0.069 (0.096)	0.040	–	–
Crowding degree	0.122 (0.092)	0.131 (0.096)	0.141 (0.097)	0.127 (0.099)	0.100 (0.100)	0.108 (0.100)	0.071	–	–
Weather and climate	0.232 (0.106)	0.127 (0.105)	0.133 (0.101)	0.226 (0.102)	0.167 (0.101)	0.144 (0.097)	0.105	–	–
Accommodation environment	0.055 (0.104)	0.043 (0.106)	0.049 (0.096)	0.065 (0.093)	0.069 (0.105)	0.059 (0.096)	0.034	1.181E-4	2.153E-6
Tour mode	0.050 (0.097)	0.054 (0.093)	0.059 (0.098)	0.065 (0.107)	0.064 (0.094)	0.052 (0.105)	0.034	1.059E-4	1.841E-6
Tourism cost	0.047 (0.108)	0.054 (0.105)	0.059 (0.094)	0.078 (0.103)	0.086 (0.098)	0.055 (0.098)	0.038	–	–
Service experience	0.092 (0.098)	0.150 (0.103)	0.091 (0.101)	0.069 (0.095)	0.076 (0.100)	0.060 (0.096)	0.054	–	–
Tour value	0.080 (0.100)	0.072 (0.105)	0.099 (0.099)	0.078 (0.098)	0.090 (0.098)	0.247 (0.101)	0.067	–	–
Natural scenery	0.169 (0.100)	0.172 (0.098)	0.145 (0.105)	0.115 (0.091)	0.112 (0.099)	0.097 (0.105)	0.081	–	–
s	0.100	0.100	0.100	0.100	0.100	0.100	–	–	–
d	3.540E-4	1.962E-4	1.389E-4	2.328E-4	1.054E-4	3.268E-4	–	–	–
exp (log)	1.050E-5	4.366E-6	2.723E-6	5.737E-6	1.830E-6	9.360E-6	–	–	–

management of scenic spots. It offers optimization strategies to provide decision-making reference for scenic-area managers for the improvement of the quality of scenic spots. The study further expands the economic advantages of tourism, takes the lead role, promotes the upgrading and development of the "demand quality" of tourism and aids in the transformation and upgrading of the tourism industry. Therefore, the contribution of this work is of great significance. This study puts forward the following development suggestions to achieve the above objectives effectively.

(1) On the basis of the analysis framework of scenic-spot portraits in this study, we can describe Jiuzhaigou, from the perspective of tourists' expression, as a scenic spot famous for its natural scenery and high-tourism value. However, given the impact of natural disasters, its infrastructure is still in the process of further improvement. As such, "travel congestion," "unreasonable tourism route planning" and other problems exist. On this basis, the infrastructure provides reference for tourists to select scenic spots.

(2) The ecological environment survey, monitoring and site visit. Descriptive statistical analysis shows that an earthquake occurred in the northwest corner of the core area of the Jiuzhaigou National Nature Reserve on August 8, 2017. This disaster has left a considerable impact on the area. The managers of scenic spots should respond to the Government actively and evaluate the reconstruction of Jiuzhaigou's scenic area.

(3) Strengthen the management of scenic spots, successfully evaluate scenic spots and formulate a scheme to manage the flow of people. The LDA analysis results reveal that tourists are dissatisfied with the "accommodation environment" and "congestion degree" of Jiuzhaigou. The current epidemic is in its normalization stage. Hence, determining how to control the flow of people is crucial to ensure tourists' safety.

(4) The augmented model analysis further verifies the tourists' attention to tourism services, such as "accommodation environment" and "congestion degree." Scenic-area managers should actively respond to the diversified and personalized needs of tourists, such as online personalized push and regular marketing.

(5) Strengthen infrastructure and information technology. The results of the sentiment analysis reveal that Jiuzhaigou attracts tourists through its scenic-natural landscape. However, its infrastructure construction degree is insufficient. The scenic area can further strengthen information technology construction, including an intelligent navigation system and Wireless Fidelity (WiFi) coverage so that tourists have a better experience in the tourism process.

5. Conclusion

Under the influence of the rapid development of the Internet, the demand for tourism information has brought about an explosive increase in tourism network information. Tourists have had increasing difficulty obtaining valuable information for such a large scale of data and excess information. In addition, the tourism strategies and online comments provided by various online tourism communities exacerbate the difficulty of potential tourists in obtaining the image of scenic spots quickly and accurately and finding scenic spots that meet their needs. At the same time, scenic-area managers struggle to understand the connotation of tourists' expression from the massive online reviews and provide targeted recommendation strategies for potential tourists. In the era of information explosion, however, information hunger seems puzzling.

Through the perspective of tourists' expression, this study proposes an LDA machine learning method based on the fusion of PHFA. Using text mining, sentiment analysis and online tourism reviews, this study analyzes the destination image of scenic spots and determines the advantages and disadvantages in the development process of scenic spots. To examine this issue further, Jiuzhaigou is taken as an example for empirical analysis. The main contributions of this study are as follows:

(1) In terms of data selection, this research uses online reviews of tourists as data for analysis to build a text-mining model for online tourism reviews. The results can reflect the preferences and needs of tourists in real time and intuitively.

(2) This study solves the problem of information overload, which causes difficulty for scenic-spot managers to determine the advantages and disadvantages in the development process of scenic spots. Moreover, it creates challenged for tourists to obtain the destination image of scenic spots quickly and accurately and find the scenic spots that meet their needs. This work provides a research paradigm for other similar studies.

(3) This study proposes a text-mining model that can reduce the uncertainty in language evaluation and provide support for the development of subsequent models. The LDA model is combined with the PHFA. The former is optimized using the TF-IDF to obtain a more accurate semantic distribution. The latter, as a new method, has high-scientific value and deserves further research and application.

(4) Overall, at the theoretical level, this study enriches the current research related to the image of scenic destinations from the perspective of tourists' reviews. It also provides a paradigm for other similar types of research. At the managerial level, this study uses an optimization model to analyze the image of Jiuzhaigou more accurately. It solves the problems regarding users' inability to obtain accurate information about the scenic area and scenic-area managers' difficulty in coping with information overload. Moreover, it provides a better understanding of the image of Jiuzhaigou from both tourists and scenic-area managers. The study also makes suggestions for the management and development of the scenic area from the perspectives of tourists and scenic-area managers.

This study has some limitations, which can be further expanded in the future. First, the emotional score calculation method based on the emotion dictionary needs a considerable amount of manpower, as expanding this dictionary involves the subjectivity of human intervention. Second, the theme determined by this study comes from specific time points and comments. In the future, we can develop an online-comment-text analysis method based on machine learning to analyze tourists' emotions dynamically and improve the satisfaction of tourists.

References

Almars, A., Li, X., Zhao, X., Ibrahim, I.A., Yuan, W.W. and Li, B.H. (2017), "Structured sentiment analysis", in Cong, G., Peng, W.C., Zhang, W.E., Li, C. and Sun, A. (Eds), *Advanced Data Mining and Applications, Adma 2017*, pp. 695-707.

Alrawadieh, Z., Alrawadieh, Z. and Kozak, M. (2019), "Exploring the impact of tourist harassment on destination image, tourist expenditure, and destination loyalty", *Tourism Management*, Vol. 73, pp. 13-20.

Bahrainian, S.A. and Dengel, A. (2013), "Sentiment analysis using sentiment features", *Web Intelligence & Intelligent Agent Technologies*, IEEE/WIC/ACM International Joint Conferences on ACM.

Baloglu, S. and Mangaloglu, M. (2001), "Tourism destination images of Turkey, Egypt, Greece, and Italy as perceived by US-based tour operators and travel agents", *Tourism Management*, Vol. 22, pp. 1-9.

Baloglu, S. and McCleary, K.W. (1999), "A model of destination image formation", *Annals of Tourism Research*, Vol. 26, pp. 868-897.

Beerli, A. and Martin, J.D. (2004), "Factors influencing destination image", *Annals of Tourism Research*, Vol. 31, pp. 657-681.

Bigne, E., Zanfardini, M. and Andreu, L. (2020), "How online reviews of destination responsibility influence tourists' evaluations: an exploratory study of mountain tourism", *Journal of Sustainable Tourism*, Vol. 28, pp. 686-704.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2002), "Latent Dirichlet allocation", in Dietterich, T.G., Becker, S. and Ghahramani, Z. (Eds), *Advances in Neural Information Processing Systems 14*, Vols 1 and 2, pp. 601-608.

Blood, D.J. and Phillips, P.C.B. (1995), "Recession headline news, consumer sentiment, the state OF the economy and presidential popularity - a time-series analysis 1989-1993", *International Journal of Public Opinion Research*, Vol. 7, pp. 1-22.

Chan, M.C., Luo, W.D. and Liu, J.N.K. (2008), "An integrated distributed system for web news retrieval", in *World Scientific Book Chapters*, Vol. 13, pp. 147-154.

Chen, Y., Zhang, H., Liu, R., Ye, Z. and Lin, J. (2019), "Experimental explorations on short text topic mining between LDA and NMF based Schemes", *Knowledge-Based Systems*, Vol. 163, pp. 1-13.

Cheng, J., Hu, S., Zhang, Y., Qi, X. and Jin, S. (2016), "Tourism destination of rural residents based on rough set and analytic hierarchy process method", *Scientia Geographica Sinica*, Vol. 36, pp. 1885-1893.

Chew, E. and Jahari, S. A. (2014), "Destination image as a mediator between perceived risks and revisit intention: a case of post-disaster japan", *Tourism Management*, Vol. 40, pp. 382-393.

Colombino, A. (2009), "Multiculturalism and time in Trieste: place-marketing images and residents' perceptions of a multicultural city", *Social and Cultural Geography*, Vol. 10, pp. 279-297.

Dai, Z.H., Dai, W., Liu, Z.H., Rao, F.Y., Chen, H.J., Zhang, G.P., Ding, Y.D. and Liu, J.Y. (2019), "Multi-task multi-head attention memory network for fine-grained sentiment analysis", in Tang, J., Kan, M.Y., Zhao, D., Li, S. and Zan, H. (Eds), *Natural Language Processing and Chinese Computing*, pp. 609-620.

Deng, G.P., Wang, L.Z., Tang, Y., Pan, Y.D., Cao, Y., Zhu, Z.F., Wang, B.X., Wang, Q.X. and Pang, W.T. (2020), "Understanding human and nature interaction outcomes for sustaining tourist destinations: an example of Jiuzhaigou Nature Reserve, China", *Aquatic Ecosystem Health and Management*, Vol. 23, pp. 373-384.

Fang, M. and Qu, Y. (2018), "Research on the concern degree of online bed and breakfast reputation evaluation based on text review", *Areal Research and Development*, Vol. 37, pp. 123-127.

Fulse, S., Sugandhi, R. and Mahajan, A. (2017), *A Survey on Multimodal Sentiment Analysis*, Esrsa Publications, Vol. 65, pp. 3-14.

Gruninger, J. (2007). "Non-negative factorization of non-negative matrices - art. no. 67480H", in Bruzzone, L. (Ed.), *Image and Signal Processing for Remote Sensing Xiii*, p. H7480.

Guo, X.X., Pesonen, J. and Komppula, R. (2020), "Comparing online travel review platforms as destination image information agents", *Information Technology and Tourism*, Vol. 36, pp. 85-93.

Guo, X., Pesonen, J. and Komppula, R. (2021), "Comparing online travel review platforms as destination image information agents", *Information Technology and Tourism*, Vol. 23, pp. 159-187.

Haselmayer, M. and Jenny, M. (2017), "Sentiment analysis of political communication: combining a dictionary approach with crowdcoding", *Quality and Quantity*, Vol. 51, pp. 2623-2646.

Hou, Z.Q. (2017), "Comparison study of advertising videos of tourism city image-based on the content analysis method", *Eurasia Journal of Mathematics Science and Technology Education*, Vol. 13, pp. 6837-6846.

Jarchi, D. and Boostani, R. (2006). "A new weighted LDA method in comparison to some versions of LDA", in Ardil, C. (Ed), *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 18, pp. 233-238.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. (2019), "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Vol. 78, pp. 15169-15211.

Jiang, J.G., Lu, Y., Yu, M., Li, G., Liu, C., Huang, W.Q. and Zhang, F.T. (2018), "Sentiment embedded semantic space for more accurate sentiment analysis", in Liu, W., Giunchiglia, F. and Yang, B. (Eds), *Knowledge Science, Engineering and Management, Ksem 2018, Pt Ii*, pp. 221-231.

Kuo, T.S., Huang, K.C. and Nguyen, P.H. (2019), "Adoption OF mobile applications for identifying tourism destinations BY travellers: an integrative approach", *Journal of Business Economics and Management*, Vol. 20, pp. 860-877.

Ladhari, R. and Souiden, N. (2020), "The role of mega-sports event experience and host city experience in explaining enjoyment, city image, and behavioral intentions", *Journal of Travel and Tourism Marketing*, Vol. 37, pp. 460-478.

Lee, S.W. and Xue, K. (2020), "A model of destination loyalty: integrating destination image and sustainable tourism", *Asia Pacific Journal of Tourism Research*, Vol. 25, pp. 393-408.

Li, Q.C., Lin, S., Dong, Z.H. and Ieee (2008), *Research of Web Information Mining by Using Crawler Techniques*, IEEE.

Liang, X., Liu, Z. and Liu, P. (2018), "Pythagorean uncertain linguistic TODIM method based on generalized Choquet integral and its application", *Control and Decision*, Vol. 33, pp. 1303-1311.

Lin, M., Zhan, Q., Xu, Z. and Chen, R. (2018), "Group decision making with probabilistic hesitant multiplicative preference relations based on consistency and consensus", *IEEE Access*, Vol. 6, pp. 63329-63344.

Lin, M., Zhan, Q. and Xu, Z. (2020), "Decision making with probabilistic hesitant fuzzy information based on multiplicative consistency", *International Journal of Intelligent Systems*, Vol. 35, pp. 1233-1261.

Lin, Z., Huang, C. and Lin, M. (2021), "Probabilistic hesitant fuzzy methods for prioritizing distributed stream processing frameworks for IoT applications", *Mathematical Problems in Engineering*, Vol. 24, pp. 104-123.

Mariani, M. and Borghi, M. (2021), "Are environmental-related online reviews more helpful? A big data analytics approach", *International Journal of Contemporary Hospitality Management*, Vol. 10, pp. 21-46.

Moon, H. and Han, H. (2019), "Tourist experience quality and loyalty to an island destination: the moderating impact of destination image", *Journal of Travel and Tourism Marketing*, Vol. 36, pp. 43-59.

Nadeem, S., Raza, M., Kayani, N., Aziz, A. and Nayab, D. (2018), "Examining cross-cultural compatibility of high performance work practices", *International Business Review*, Vol. 27, pp. 563-583.

Palau-Saumell, R., Forgas-Coll, S., Amaya-Molinar, C.M. and Sanchez-Garcia, J. (2016), "Examining how country image influences destination image in a behavioral intentions model: the cases of Lloret De Mar (Spain) and Cancun (Mexico)", *Journal of Travel and Tourism Marketing*, Vol. 33, pp. 949-965.

Pestana, M.H., Parreira, A. and Moutinho, L. (2020), "Motivations, emotions and satisfaction: the keys to a tourism destination choice", *Journal of Destination Marketing and Management*, Vol. 16, pp. 34-87.

Pike, S. (2002), "Destination image analysis - a review of 142 papers from 1973 to 2000", *Tourism Management*, Vol. 23, pp. 541-549.

Ren, P.J., Xu, Z.S., Wang, X.X. and Zeng, X.J. (2021), "Group decision making with hesitant fuzzy linguistic preference relations based on modified extent measurement", *Expert Systems with Applications*, Vol. 171, pp. 41-57.

Shareef, H., Ibrahim, A.A. and Mutlag, A.H. (2015), "Lightning search algorithm", *Applied Soft Computing*, Vol. 36, pp. 315-333.

Song, Y., Jeong, S. and Kim, H. (2017), "Semi-automatic construction of a named entity dictionary for entity-based sentiment analysis in social media", *Multimedia Tools and Applications*, Vol. 76, pp. 11319-11329.

Sun, J.X., Wang, X.R. and Ma, T. (2018), "De-localization and re-localization of ethnic cultures: what is the role of tourism?", *Asia Pacific Journal of Tourism Research*, Vol. 23, pp. 1034-1046.

Suo, Z., Liang, L., Su, X., Mai, Z. and Zhang, J. (2019), "Visitor's experience and tourism destination image in Kaifeng city:based on Internet comment text", *Areal Research and Development*, Vol. 38, pp. 102-105.

Tunali, V., Bilgin, T. and Camurcu, A. (2016), "An improved clustering algorithm for text mining: multi-cluster spherical k-means", *International Arab Journal of Information Technology*, Vol. 13, pp. 12-19.

Wan, H., Zhang, M. and Song, H. (2017), "Research on tourists' satisfaction with the major scenic spots in Baoji city", *Areal Research and Development*, Vol. 36, pp. 99-102.

Wang, Y., Zhang, Z. and Huo, W. (2016), "Research on aviation unsafe incidents classification with improved tf-idf algorithm", *Modern Physics Letters B*, Vol. 30, pp. 16-44.

Wei, W., Liu, Y.P. and Wei, L.R. (2020), "Feature-level sentiment analysis based on rules and fine-grained domain ontology", *Knowledge Organization*, Vol. 47, pp. 105-121.

Wu, G. and Liang, L.F. (2020), "Examining the effect of potential tourists' wine product involvement on wine tourism destination image and travel intention", *Current Issues in Tourism*, Vol. 24, pp. 2278-2293.

Wu, Y.H., Ding, Y.X., Wang, X.L. and Xu, J. (2010), "A comparative study of topic models for topic clustering of Chinese web news", in Hang, Y., Desheng, W. and Sandhu, P.S. (Eds), *Proceedings of 2010 3rd Ieee International Conference on Computer Science and Information Technology*, pp. 236-240.

Xi, Y., Yu, H., Yao, Y., Peng, K., Wang, Y. and Chen, R. (2020), "Post-traumatic stress disorder and the role of resilience, social support, anxiety and depression after the Jiuzhaigou earthquake: a structural equation model", *Asian Journal of Psychiatry*, Vol. 49, pp. 1-17.

Xiao, X., Fang, C.Y. and Lin, H. (2020), "Characterizing tourism destination image using photos' visual content", *ISPRS International Journal of Geo-Information*, Vol. 9, pp. 73-89.

Yin, S.M., Niu, Y.G. and Li, W. (2015), "Advertising information content of web forums and message boards official posts in China National 5A rated tourism attractions", in Xu, J., Nickel, S., Machado, V.C. and Hajiyev, A. (Eds), *Proceedings of the Ninth International Conference on Management Science and Engineering Management*, pp. 459-470.

Young-Seok, S. and Kim, H.-b. (2016), "A study of destination image and measurement using text mining", *Journal of Tourism Sciences*, Vol. 40, pp. 221-245.

Yuan, Y.F., Du, J.P., Lee, J. and IEEE (2016), "Tourism activity recognition and discovery based ON improved LDA model", *Proceedings of 2016 4th Ieee International Conference on Cloud Computing and Intelligence Systems*, pp. 447-455.

Zhang, S., Xu, Z.S. and He, Y. (2017), "Operations and integrations of probabilistic hesitant fuzzy information in decision making", *Information Fusion*, Vol. 38, pp. 1-11.

Zhao, J., Liu, K. and Xu, L.H. (2016), "Sentiment analysis: mining opinions, sentiments, and emotions", *Computational Linguistics*, Vol. 42, pp. 595-598.

Zhong, S.E., Zhang, J. and Sun, J. (2015), “The diversified development of historic towns based on the integrated evaluation system of heritage conservation and tourism development in Jiangsu”, *Geographical Research*, Vol. 34, pp. 1380-1393.

Corresponding author
Tao Tong can be contacted at: tongtao2021@qq.com