

Research on Online Review of Ice and Snow Sports Based on Text Mining

Yingkun Zhang^{1,*}, Yuliang Chen¹, Zijian Mo¹, Xiaoxue Yu², Luhongyu Lin³

¹ Shijiazhuang Tiedao University, Shijiazhuang 050000, China;

² Nanjing University of Technology, Nanjing 210000, China;

³ Central South University for Nationalities, Wuhan 430000, China.

* Corresponding Author Email: 1540838594@qq.com

Abstract. Ice and snow economy, with the development of times and economy and the heat of 2022 Winter Olympic Games, ice and snow sports have gradually become a trend. However, domestic and foreign scholars to this emerging economy research is less, the level is shallow. This paper takes Ctrip platform as the data source, grabs the online review text of ice and snow sports, processes the data, quantifies the comment attribute of consumers by TF-IDF method, then uses LDA model to explore its potential theme, and finally constructs emotion classification model to analyze consumers' emotion. The purpose is to understand their emotional bias and provide reference suggestions for national decision-making and ice and snow economic development.

Keywords: Ice and snow economy; Online comments; TF-IDF; LDA thematic model; Emotion classification model.

1. Introduction

1.1. Study Background

With the development of the economy, people have begun to get involved in ice and snow sports. Skating rinks and ski resorts have become places for people to relax and decompress on weekends. The infrastructure related to the Winter Olympics keeps pace with the progress of China's infrastructure, and these infrastructures will also benefit Beijing and Zhangjiakou after the Winter Olympics. In 2016, when General Secretary Xi participated in the deliberation of the representative delegation of Heilongjiang Province during the National Two Sessions, he put forward an important statement that "ice and snow are also golden mountains and silver mountains." On November 2, 2016, the General Administration of Sports of the People's Republic of China held a press conference to officially release the Development Plan for Ice and Snow Sports (2016-2025). Although the epidemic situation has a great impact on the ice and snow market, with the gradual control of the epidemic situation, China's ice and snow industry has begun to recover. Ice and snow economy plays a very strong leading role in the development of entertainment industry, transportation industry, catering industry and ice and snow sports industry, so as to promote the sustainable development of ice and snow economy^[1].

1.2. Research status at home and abroad

This paper mainly draws lessons from many domestic text mining methods: Zhang Hongbin (2020) takes machine learning as the research method, selects predictive variables based on literature and text mining, and uses decision tree, Logistic regression and other models. The research results provide suggestions for machine learning in the field of financial management [2]; Manju Venugopalan (2022) proposed that LDA model is an unsupervised aspect word extraction method, which is a guided Dirichlet allocation model, and uses the minimum aspect seed word in each aspect category to guide the model to identify the hidden topics of interest to consumers [3]; Khalid Y. Aram (2022) proposed a knapsack maximum edge feature selection filtering method based on SVM to select relevant independent features for support vector machine classifier in a cost-sensitive way. The results of this

study show that this model has good competitive performance as an effective cost-sensitive embedded feature selection method [4].

At present, ice and snow economy has become a new economy, and scholars at home and abroad have little research on it, and domestic text mining generally focuses on policy analysis, prediction and online commodity review (Yang Runtian [5] ,2019;Liu Jianming [6], 2022).Based on the previous research, this paper uses Houyi collector to collect the comments and related information of ice and snow sports, and quantifies the comments attributes of consumers through TF-IDF method, uses LDA model to mine its deep theme, and finally analyzes the consumers 'sentiment, constructs the sentiment classification model, and compares it to get a better model to promote the development of ice and snow economy.

2. Research Framework and Methodology

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

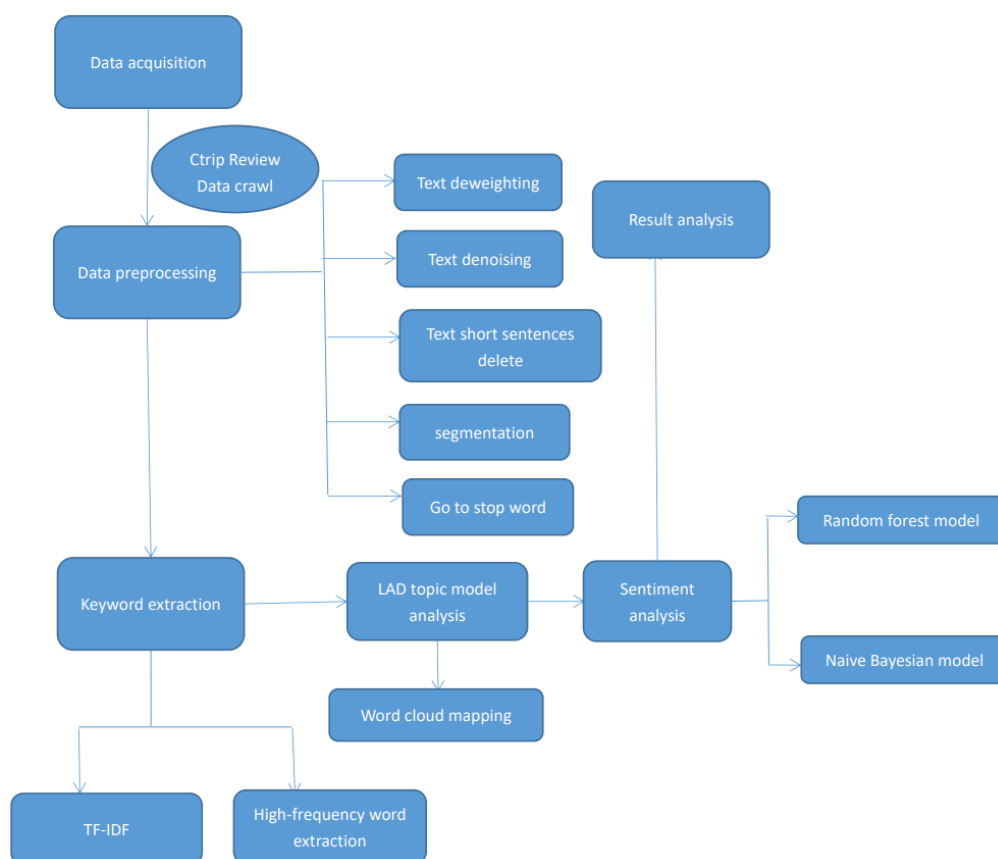


Figure 1. Overall Emotional Tendency Chart of Ice and Snow Consumer Reviews

2.1. Topic Analysis Based on LDA Model

LDA has two meanings, one is linear discriminant analysis and the other is probabilistic topic model, and the Latent Dirichlet Distribution (LDA) used in this paper ^[7].The purpose is to identify the topic, first from the **document** containing many words to extract the relevant topics, and then extract the general vocabulary, loop iteration until all the words in the document.The flow chart is shown in Figure 2 below:

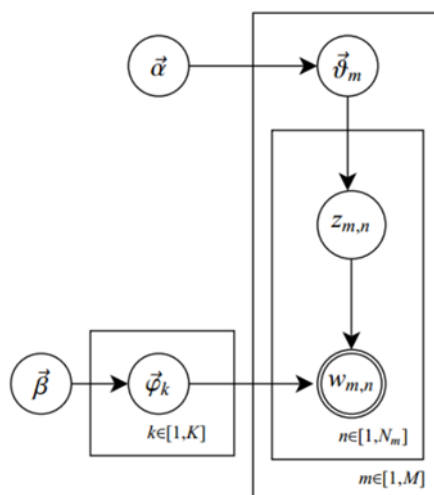


Figure 2. Schematic diagram of LDA topic model

Ctrip ski review data stand in the perspective of consumers, from this angle to dig the impact of ice and snow economic development focus can objectively find out the shortcomings and constantly improve. This section uses LDA theme model to analyze the advantages and disadvantages of ice and snow economy in its development, and puts forward corresponding keywords and weights to better grasp the development prospect of ice and snow economy, and gives corresponding strategies to promote its development.

2.2. Sentiment Analysis Based Global Review Analysis

The first step in applying machine learning to sentiment analysis is to define the problem, that is, to determine whether sentiment analysis is a regression problem, a classification problem or a clustering problem. In the process of semantic mining and tendency analysis of collected comment text data, user's comments include positive, negative and fair, so it can be defined as multi-classification problem. The second step is to consider ways to use classifiers.

This paper chooses Python software and uses SnowNLP module, which is one of the relatively few natural language processing libraries for Chinese design. It is convenient and easy to use for Chinese text correlation analysis and processing. Its functions are similar to other natural language processing libraries, such as part-of-speech tagging, word segmentation, sentiment analysis, etc., which can complete the content of this paper.

3. Data acquisition and processing

3.1. Data Acquisition

This paper mainly uses Houyi collector to crawl and analyze the ice and snow sports with more consumer comment data on Ctrip platform, and obtains 35895 pieces of data in total. The method comprises the following steps of: opening a page of a corresponding ice and snow sports scenic spot, copying a website of the corresponding ice and snow sports scenic spot to a collector, extracting comment data in a list, setting automatic page-turning collection and paging buttons, and running to obtain a crawling result.

3.2. Data preprocessing

Because the original data obtained in this paper has many problems, such as data repetition, data redundancy, no practical significance and so on, which will affect the analysis results, it is particularly

important to carry out data preprocessing, data cleaning, text segmentation, stop words removal to improve the quality of data, and provide more favorable support for subsequent analysis.

3.2.1 Data cleaning

① In order to improve their own benefits, some e-commerce platforms usually set up some automatic programs to automatically generate favorable comments for consumers who have not been evaluated for a long time. At the same time, some consumers are forced to evaluate in order to enjoy some preferential treatment, copying others' comments at will or commenting on short text sentences without substantive information. The large amount of duplicate content is not intended, so this article should remove the invalid data.

② Text denoising is to delete the blank space, expression, number, English character and other special symbols in the text that have no practical significance for analysis.

3.2.2 Text segmentation

A participle is a word that is broken up into several words. Due to the existence of a large number of Internet catchwords and some proper nouns (such as "ice mound mound" and "snow melt ") that are easy to cause ambiguity after being segmented, this article uses the jieba in python to create a custom dictionary containing the keyword skiing comment text.

3.2.3 De-stop word

After word segmentation, words such as "de,""le,""ah,""ba" and "this" will appear in the text, which are meaningless to the analysis of ice and snow economy. The high frequency of these words will increase the complexity of the language and affect the subsequent word frequency statistics, so stop word filtering is needed.

3.3. Keyword extraction

3.3.1 The concept and algorithm of TF-IDF

In this paper, TF-IDF algorithm will be selected as the keyword extraction method of Ctrip review data [8]. The default hypothesis is that words with high frequency are relatively important. TF-IDF consists of two parts:

TF refers to the word frequency, which is usually normalized, that is, the word frequency divided by the total number of words in the article, so that it does not bias towards long files. However, some common words are not very helpful to the topic, and it is possible that t_i words that appear less frequently can represent the theme of the article, the TF value is not appropriate at this time. Hence introduce IDF , emphasize the meaning of the items in the set as a whole, and balance the setting of weights.

$$IDF_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (1)$$

$|D|$ is that numb of files in the corpus, $|\{j:t_i \in d_j\}|$ is the number of files containing a word. If the word is not in the corpus, the denominator will be 0. Therefore, the following formula is used for calculation:

$$IDF_i = \log \frac{|D|}{|\{j:t_i \in d_j\}| + 1} \quad (2)$$

Then, the value of TF-IDF is:

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i \quad (3)$$

The core idea of the algorithm is that the high frequency in the current document and the low frequency in the document set indicate that the word is very important and has good ability to

distinguish categories. Furthermore, the high-frequency words can be extracted and sorted according to the TF-IDF weight.

3.3.2 High-frequency word extraction

High-frequency words refer to those that appear more frequently in the review data and have reference value for service attributes. In this paper, we do a simple word frequency analysis of consumer reviews, and then judge their concerns, and then process the effective reviews into word segmentation and stop words to form a corpus.

According to the summary results, the six keywords with the highest frequency are "skiing," "ski resort," "good," "coach," "ski resort" and "ski path," which indicates that consumers have a high pursuit of ski resort facilities and environment. The ice and snow economy has been further developed. Therefore, in order to keep the heat of ice and snow economy unabated, the quality of ski-related services or goods such as ski resorts should be continuously improved.

4. Analysis of Ice and Snow Sports Reviews Based on LDA Topic Model

This paper uses LdaModel of gensim module in python as theme model, preprocesses the evaluation data of ski-related services by consumers on Ctrip platform, creates custom dictionary, removes stop words, then processes word segmentation, and then imports the data after word segmentation into the model. The concept of degree of confusion is used in the selection of topics.

The smaller the uncertainty of the trained model about which topic it belongs to, the better the quality of the model. After 1000 iterations, it is found that when the number of topics is set to 2, the confusion is the least, so the portable comment data is divided into 2 topics, and 20 keywords are extracted from each topic.

Table 1. Analysis of LDA Theme Model of Ski Resort Ice and Snow Sports Scenic Spot Reviews

positive thematic analysis		negative theme analysis	
Coach	0.020	Queue up	0.009
Piste	0.012	Staff	0.007
Happy	0.012	Place	0.007
Convenient	0.010	Services	0.006
Suit	0.010	Experience	0.006
Service	0.009	Ctrip	0.005
Junior	0.009	Management	0.004
Children	0.008	Cable Car	0.004
Next Time	0.007	Poor	0.004
Facilities	0.007	Snow Road	0.004
Fun	0.017	Coach	0.017
Snow	0.013	Rent	0.006
Place	0.010	Staff	0.006
Worth	0.009	Scenic	0.006
Experience	0.007	Tickets	0.006
Stimulus	0.006	Queue	0.005
Fall	0.006	Hotel	0.005
Winter	0.006	Ski	0.005
First Time	0.006	Suggestion	0.004
Special	0.005	Deposit	0.004

1. In the positive theme, the keywords such as "piste," "coach," "fun," "fun," and "snow" reflect people's high requirements for ski resorts and ski equipment.

2. In the negative theme, "coach" is still a high-frequency keyword. From this analysis, it can be seen that people have higher requirements for personal safety and fun brought by sports during skiing,

and feel the "Winter Olympics Spirit" personally. The government should pay attention to this point, add luster to ice and snow services, and increase employment positions from the position of "coach" to relieve the pressure of employment; At the same time, "queue" and "staff" are also high-frequency keywords, so it is necessary to expand the scale and reduce the problems such as "queue" and "staff."

5. Analysis of Ice and Snow Consumer Preference and Decision-making Based on Review Data

5.1. WOM Emotion Score of Ice and Snow Consumers Based on Review Data

According to the classification of the word-of-mouth data, the consumer's comments on ice and snow can be divided into three categories, namely positive comments, neutral comments and negative comments; Quantitative analysis of the data can be used to obtain satisfaction scores and emotional tendencies that are positive or negative.

This paper will collect the data through the Snow-NLP(line) to score each comment, and all comments under each category score summary into a bar chart, 1 for positive emotion, 0 for neutral emotion, -1 for negative emotion.

Using the SnowNLP database, the score value of the comment data greater than 0.6 is judged as a positive comment, less than 0.4 is judged as a negative comment, and the rest is judged as a neutral comment. Some results are shown in the table above.

According to the above data results, it can be concluded that the vast majority of consumers hold a positive attitude towards ice and snow experience, are willing to try and learn ice and snow events, and are very interested in ice and snow sports. Some consumers only follow the crowd psychology, or have never tried the ice and snow project, after the first attempt according to the experience and cost gave a more pertinent evaluation, there are some consumers to tickets, facilities, experience, diet and other aspects of dissatisfaction, gave negative comments.

The histogram and analysis results obtained in this paper are presented as shown in the following figure:

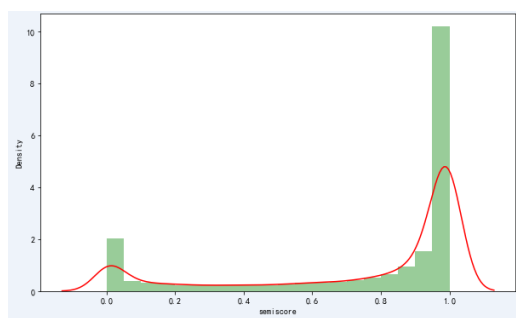


Figure 3. Overall Emotional Tendency Chart of Ice and Snow Consumer Reviews

It can be seen from the overall emotional tendency chart of comments that most consumers are very positive about the ice and snow experience, some of them give moderate comments, and a few give negative comments. Therefore, the government needs to improve facilities, tickets, services, environment and other aspects, perfect the deficiencies, and provide consumers with perfect experience, so as to obtain greater economic benefits.

5.2. Analysis of Ice and Snow Consumer Preference Based on LDA Model

5.2.1 Thematic Analysis of "Most Satisfied" Review Data of Ice and Snow Consumers

Table 2. Analysis of the "Most Satisfied" Comments on Ice and Snow Sports

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
Key words	Key words	Key words	Key words	Key words
weight	weight	weight	weight	weight
Coach 0.020	Special 0.005	Services 0.009	Stimulus 0.006	First time 0.006
Piste 0.012	Experience 0.007	Convenient 0.010	Fun 0.017	Throw 0.006
Primary 0.009	Worth 0.009	Place 0.010	Happy 0.012	Snow 0.013
Facilities 0.007	Next time 0.007	Fit 0.001	kid 0.008	Winter 0.006

According to the results of the LDA model of the "most satisfied" review data, the model divides the review data into five topics.

The key words of theme 1 include "coach," "trail," "elementary" and "facilities," which reflect consumers' intuitive feelings about ice and snow sports. One of the satisfaction points of most consumers is that the coach has a good attitude, teaches seriously, and has complete equipment and facilities. The government can improve the arrangement of site and personnel accordingly.

The words "special," "experience," "worth" and "next time" in theme 2 reflect the psychological situation of consumers. They feel good about the experience of the projects they have tried and hope to spend again next time. From this, it can be seen that the government should pay attention to the experience of consumers and provide better services.

The "service," "convenience," "place" and "suitability" in theme 3 reflect that the development of ice and snow sports has the support of government policies, which is also one of the points of consumer satisfaction. At the same time, the service attitude is good, the travel traffic is convenient, and the site is suitable for people of all levels.

The words "exciting," "fun," "happy" and "children" in Theme 4 reflect consumers' enjoyment of ice and snow sports, with some parents bringing their children to play, while others are youth groups to experience exciting projects and relax. Therefore, the government should explore and innovate in time, research new projects, new play methods, so as to keep consumers interested and enthusiastic.

In the fifth theme, "the first time," "fall," "snow" and "winter" reflect that winter is frozen thousands of miles, covered in silver, and most consumers come to the north to experience different ice and snow sports. Many southerners have never seen snow, so they are excited to see it for the first time. The unique geographical location and climate in the north form a unique scenery in winter, attracting many consumers to come to play. This promotes the development of local ice and snow sports industry and drives the development of economy. The state should attach great importance to ice and snow economy.

Based on the above five themes, it can be seen that consumers' intuitive feelings on ice and snow services, consumers' psychological situation, consumers' interests and other factors affect the development of ice and snow economy.

5.2.2 Thematic Analysis of the "Most Dissatisfied" Review Data of Ice and Snow Consumers

Table 3. Analysis of "Most Dissatisfied" Comments on Ice and Snow Sports

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
keyword weight	keyword weight	keyword weight	keyword weight	keyword weight
Staff 0.009	Expenses 0.006	Services 0.004	Deposit 0.006	Piste 0.005
Services 0.007	Queue up 0.005	Attitude 0.004	Rent 0.006	Difference 0.005
Coach 0.007	Tickets 0.004	Experience 0.017	Advice 0.006	Food 0.004
Scenic spot 0.006	Ctrip 0.004	Time 0.006	Place 0.005	Facilities 0.004

According to the results of the LDA model of the "least satisfied" review data, the model divides the review data into five topics.

The key words of theme 1 include "staff," "service," "teaching and training" and "scenic spot," which reflect consumers' intuitive experience of ice and snow service. Most consumers are

dissatisfied with the coach's poor attitude, careless teaching, imperfect equipment and other facilities, and the venue is not suitable for people of all levels. Therefore, the government should increase investment in sites, personnel and facilities.

Theme 2 of the "tolls," "queue," "tickets" and "carry Cheng" and so on reflect the consumer ice and snow sports spending expectations, tolls, tickets and other expenses feel expensive, play pressure. It can be seen that the government should consider the economic situation of consumers and control prices.

The "service," "attitude," "experience" and "time" in theme 3 reflect that consumers have higher requirements for ice and snow sports services, hope for better service attitude and better experience, so catering to consumers is the primary goal of ice and snow sports development, and meeting consumers' needs is another important factor to promote the development of ice and snow economy.

The words "deposit," "rent," "suggestion" and "place" in Theme 4 reflect that consumers do not understand the extra fees such as deposit for ice and snow sports events, are unwilling to pay, etc., and have no plan for the sequence of ice and snow events. I hope the organizer can put forward reasonable play strategies. Therefore, the organizer should reasonably determine the amount of deposit, ensure that the amount of deposit is within the acceptable range of the public, and at the same time explore and innovate, research new projects and new play methods, and stimulate consumers' interest.

In Theme 5, "snow path," "poor," "food" and "facilities" show that consumers are not satisfied with the experience of ice and snow sports events. The food is expensive, the service is poor, the personnel is crowded, the facilities are old, etc., which sometimes occur, resulting in some consumers' disappointment. Therefore, the government should pay attention to the real-time situation of ice and snow sites, change the status quo and meet the needs of consumers.

Based on the above five themes, it can be seen that the feelings of consumers have affected the development of the ice and snow economy. Therefore, the government should pay attention to the thoughts of consumers, so as to promote the development of ice and snow economy.

Through the foregoing analysis, we can get the psychological preference and demand tendency of consumers to choose ice and snow sports, so as to get the main influence of whether consumers participate in ice and snow sports. Based on the text comment data, this chapter makes the following five comment models and compares them.

5.3. Sentiment classification model based on ice and snow consumer review data

5.3.1 Preprocessing and Evaluation Index of Sentiment Classification Model

This chapter uses the ice and snow consumer comment tablet data on Ctrip platform. After preprocessing, it is substituted into two classification models, namely random forest model [10] and naive Bayesian model and [9]. In this chapter, the data with clear labels are created in each model according to two feature creation methods, four analysis results are obtained and compared, and the classification model under the optimal feature creation method is selected. The evaluation data of ice and snow sports are divided into three categories: "positive," "negative" and "neutral." In this paper, the "positive" emotion is defined as liking ice and snow sports and willing to try and experience ice and snow events, the "negative" emotion is defined as unwilling to try and experience ice and snow events, and the "neutral" emotion is defined as either experiencing or not experiencing ice and snow events, which are marked as 1, -1 and 0 respectively. Finally, there are 34327 samples. Among them, 25733 were willing to undergo physical examination, 6792 were unwilling to experience, and 1802 could be experienced or not. This article builds a classification model with python based on this data. The classification models used in the next section of this article can be loaded in the sklearn module.

After the completion of the model, this paper needs to evaluate the model. The evaluation indexes mainly include precision rate, recall rate, accuracy rate, F1 Score, etc.

The formula is as follows:

$$\text{Accuracy: } Accuracy = \frac{TP + FN}{TP + TN + FP + FN} * 100\% \quad (4)$$

$$\text{Recall Rate: } Precision = \frac{TP}{TP + FP} * 100\% \quad (5)$$

$$\text{Recall: } Recall = \frac{TP}{TP + TN} * 100\% \quad (6)$$

$$\text{F1 Score: } F_1\text{score} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

Before making each model, we should first collect and sort out the data. The consumer's evaluation data of ice and snow sports used in this paper have been pre-processed above and can be used directly here.

5.3.2 Sentiment classification model based on ice and snow consumer evaluation data[10]

① Sentiment classification model based on random forest

Random forest is a machine learning model. A general random forest has the following characteristics:

Method 1: The prediction results of feature extraction using TfidfVectorizer are as follows:

	precision	recall	f1-score	support
-1	0.65	0.53	0.58	1333
0	0.17	0.04	0.06	363
1	0.84	0.93	0.89	5170
accuracy			0.81	6866
macro avg	0.55	0.50	0.51	6866
weighted avg	0.77	0.81	0.78	6866

Figure 4. Evaluation diagram of sentiment classification model based on random forest

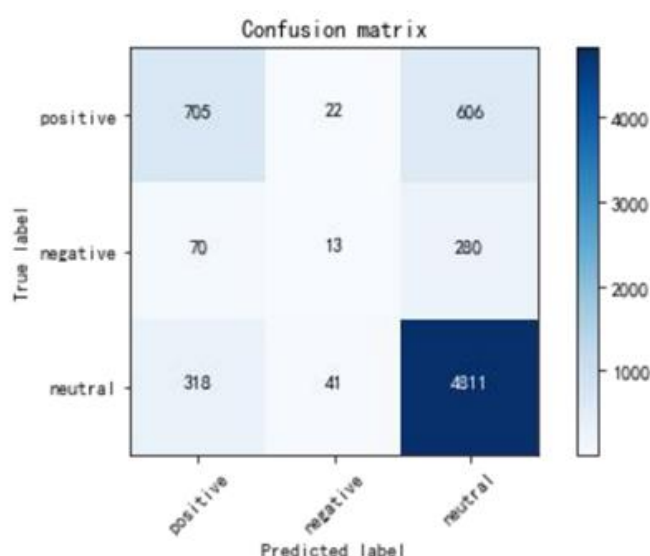


Figure 5. Confusion matrix a of classification model based on random forest

The evaluation index of the sentiment classification model based on the random forest is shown in FIG. 6. According to the confusion matrix, there are 5529 correct classifications, 350 wrong

classifications that should be negative emotions, 628 wrong classifications that should be positive emotions, and 359 wrong classifications that should be neutral emotions.

Method 2: Pretest results of feature extraction using CounterfVectorizer are as follows:

	precision	recall	f1-score	support
-1	0.60	0.56	0.58	1333
0	0.20	0.07	0.10	363
1	0.85	0.91	0.88	5170
accuracy			0.80	6866
macro avg	0.55	0.51	0.52	6866
weighted avg	0.77	0.80	0.78	6866

Figure 6. Evaluation diagram b of sentiment classification model based on random forest

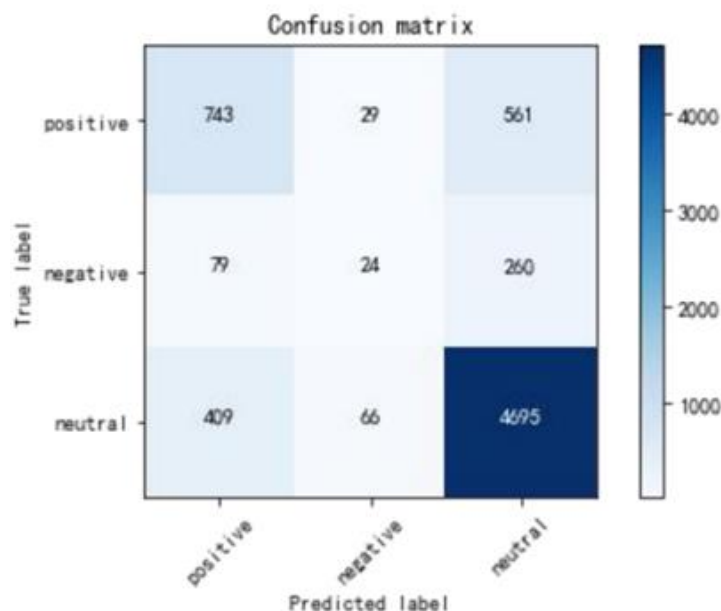


Figure 7. Confusion matrix b of classification model based on random forest

The evaluation index of the sentiment classification model based on the random forest is shown in FIG. 8. According to the confusion matrix, there are 5462 correctly classified emotions, 339 wrongly classified emotions that should be negative, 590 wrongly classified emotions that should be positive, and 475 wrongly classified emotions that should be neutral.

From the above evaluation indexes, this paper can see that the effect of classification using random forest algorithm is better, this algorithm is relatively simple, can deal with high latitude data, can balance error and has strong anti-interference ability.

② Sentiment classification model based on Naive Bayes

Method 1: The prediction results of feature extraction using TfidfVectorizer are as follows:

	precision	recall	f1-score	support
-1	0.88	0.47	0.61	1333
0	0.00	0.00	0.00	363
1	0.83	0.99	0.90	5170
accuracy			0.83	6866
macro avg	0.57	0.48	0.50	6866
weighted avg	0.79	0.83	0.80	6866

Figure 8. Evaluation diagram of sentiment classification model based on Naive Bayes

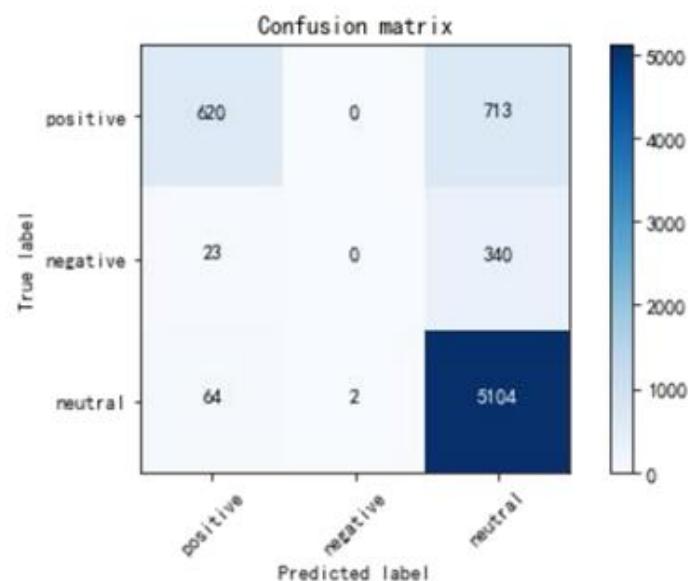


Figure 9. Naive Bayes-based sentiment classification model confusion matrix a

The accuracy of Naive Bayes sentiment classification model is 0.83. This paper only focuses on weighted average. The evaluation indicators of the classification model are shown in FIG. 10. From the confusion matrix, it can be seen that 5724 items are correctly classified, 23 items are negative emotions wrongly classified into positive emotions, 64 items are neutral emotions wrongly classified into positive emotions, 340 items are negative emotions wrongly classified into neutral emotions, and 713 items are positive emotions wrongly classified into neutral emotions.

Method 2: Pretest results of feature extraction using CounterVectorizer are as follows:

	precision	recall	f1-score	support
-1	0.72	0.69	0.70	1333
0	0.12	0.01	0.02	363
1	0.88	0.94	0.91	5170
accuracy			0.84	6866
macro avg	0.57	0.55	0.55	6866
weighted avg	0.81	0.84	0.82	6866

Figure 10. Naive Bayes-based sentiment classification model evaluation diagram b

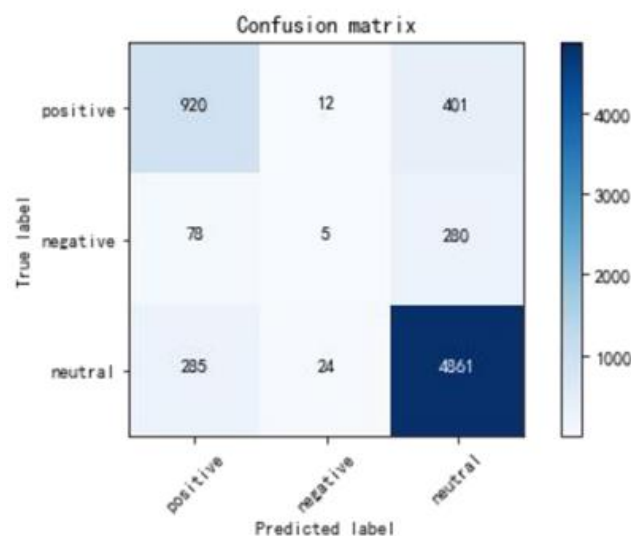


Figure 11. Confusion matrix b of sentiment classification model based on Naive Bayes

The classification model sentiment analysis is shown in Figure 12. According to the confusion matrix, there are 5786 correct classification items, 78 negative emotions are wrongly classified into positive emotions, 285 neutral emotions are wrongly classified into positive emotions, 280 negative emotions are wrongly classified into neutral emotions, and 401 positive emotions are wrongly classified into neutral emotions.

From the above evaluation indexes, we can conclude that Naive Bayes algorithm has better classification effect, stable classification efficiency, less complexity and faster query speed.

5.3.3 Comparison and analysis of the two sentiment classification models.

According to the evaluation index in the previous section, this paper can see that the sentiment classification model based on random forest and the sentiment classification model based on naive Bayes have better effect and relatively high quality. In this paper, the two classification models under different feature extraction methods are compared, and the results are as follows:

Table 4. Comparison of Efficiency

Classification model	Feature extraction mode	Accuracy	Run time
Random forest	TfidfVectorizer	0.81	3.6612s
	CounterVectorizer	0.80	7.8570s
Naive Bayes	TfidfVectorizer	0.83	0.0289s
	CounterVectorizer	0.84	0.0189s

From the model data, it can be seen that the TfidfVectorizer feature extraction mode based on the random forest classification model has higher accuracy and shorter travel time; In the CounterVectorizer feature extraction mode based on the naive Bayesian model, the accuracy rate is higher, and the running time is shorter. However, the sentiment classification model based on naive Bayes model has the best classification effect on ice and snow sports comment data. Consumers can use this classification model to decide whether to participate in ice and snow sports, thus driving the development of ice and snow sports.

6. Countermeasures and Suggestions

First, formulate the development plan of special land for ice and snow sports, and give strong policy support to the implementation project of ice and snow site construction. Strive to meet the

initial demand of consumers to participate in ice and snow sports, and cultivate potential skiing consumers. On the other hand, through the holding of "skiing events," to create a gathering highland for ice and snow events. Policy support has been issued to support the national ice and snow industry and encourage entrepreneurs to start businesses.

7. Conclusion

According to the key words obtained by LDA theme analysis method, it can be seen that the main problems in the development of China's ice and snow economy are as follows: The supply of ice and snow sites is insufficient, and the transportation infrastructure support capacity is poor. And this happens to be the primary problem for the smooth development of ice and snow sports. In order to stimulate the development of ice and snow economy, we should increase investment in the construction of both.

According to the research on online reviews of ice and snow sports, most consumers feel positive about ice and snow sports, but some consumers also give pertinent or negative comments. The government can take this as a breakthrough to improve the facilities, tickets, services, environment and other aspects that consumers are dissatisfied with, so as to give consumers a more perfect experience and obtain greater economic benefits.

In addition, there is also a shortage of ice and snow service personnel in our country, which makes it difficult to provide supplies for the large ice and snow market. With the rapid development of ice and snow industry, our country should pay attention to increasing the investment in training ice and snow related professionals, so as to provide necessary conditions and policy support.

References

- [1] Zhang Guihai. Create a new model for ice and snow industry development [J]. *Struggle*, 2021 (02):48.
- [2] Zhang Hongbin, Guo Meng. Machine Learning and Financial Forecasting--Empirical Research on the Application of Early-warning System of Listed Companies in China [J]. *Finance Quarterly*, 2020, 14 (04):135-154.
- [3] Venugopalan Manju, Gupta Deepa. An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis [J]. *Knowledge-Based Systems*, 2022, 246.
- [4] Aram Khalid Y., Lam Sarah S., Khasawneh Mohammad T.. Linear Cost-sensitive Max-margin Embedded Feature Selection for SVM [J]. *Expert Systems With Applications*, 2022, 197.
- [5] Yang Runtian, Xu Tengda. The Development Scale of Chongli Ski Tourism Industry under the Background of Winter Olympic Games--Based on the Perspective of Economic Prediction [J]. *Shenyang Institute of Physical Education*, 2019, 38 (06):1-7.
- [6] Liu J M, Mi J, Jin N J, Liu J Z. Evolution Characteristics and Prospect of China 's Ice and Snow Sports Industry Policy [J]. *Sports Culture Guide*, 2022 (02):1-7.
- [7] Wang XW, Zhang L, Huang B, Wei Y N. Construction and Empirical Study of Consumer Topic Map of Microblog Based on LDA--Taking "Egyptian Airlines Disaster" as an Example [J]. *Data Analysis and Knowledge Discovery*, 2020, 410:47-57.
- [8] Yang Kaiyan. Research on keyword automatic extraction algorithm based on improved TFIDF [D]. *University of Chicago*, 2015.
- [9] Shi Yixue, Yu Zhengtao, Xiang Yan, Zhang Yafei. China-Vietnam Cross-language Sentiment Classification Model Incorporating Topic Features [J]. *Chinese Journal of Information Science*, 2022, 36 (03):36-44.
- [10] Du Keke, Zhang Yue, Zhao Kai. Personal credit evaluation model combining similarity measure and random forest [J]. *Journal of Chongqing University of Industry and Commerce (Natural Science Edition)*, 2022, 39(03):54- 60. DOI:10.16055/j.issn.167 2-058X.2022.0003.008.
- [11] Wang Guocai. Research and Application of Naive Bayes Classifier [D]. *Chongqing Jiaotong University*, 2010.