Research on Big Data Analysis of Passenger-Oriented Railway Service Quality Based on LDA Model

Haohuan Yuan[1](✉), Hao Ding[2], and Pengcheng Lv[3]

[1] School of Mathematics, Hefei University of Technology, Hefei, Anhui, China
duoerxs@163.com

[2] School of Computer Science and Technology, Harbin Institute of Technology, Weihai, Shandong, China

[3] College of Shipbuilding Engineering, Harbin Engineering University, Harbin, Heilongjiang, China

Abstract. Because of the lack of a reasonable evaluation and feedback system in China's railway transportation sector and the inability to make timely improvement according to the needs of passengers, this paper presents a passenger-oriented railway service quality big data analysis and research based on Latent Dirichlet Allocation (LDA) model. This method is of positive significance for railway departments to establish their evaluation and feedback system and opinion analysis. This paper excavates the needs of passengers in railway transportation and uses the Term Frequency-Inverse Document Frequency (TF-IDF) method to calculate text feature words and conducts semantic network analysis on them. Based on the above, the LDA topic model is used for modeling, and the best number of topics of comment data is determined by calculating the model confusion, and the LDA topic model is trained according to the best number of topics. Through the analysis of the model, the existing problems of railway traffic service based on the different demands of passengers are explored, and the emotional tendencies of current railway passengers are summarized based on the sentiment analysis algorithm.

Keywords: Latent Dirichlet Allocation · railway service quality · sentiment analysis · data mining

1 Introduction

As one of the transportation departments in China, the railway sector plays a significant role in People's Daily life. However, compared with other service places, railway departments do not have a set of evaluation and feedback systems to timely obtain passengers' suggestions and demands for railway traffic services [1, 2]. Therefore, even if passengers are not satisfied with the service quality of railway departments, their demands cannot be effectively fed back due to the lack of a reasonable evaluation and feedback system. As a result, the railway department cannot make timely and practical adjustments and solutions to the needs and suggestions of passengers. As a direct result of this phenomenon,

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 1279–1293, 2023.
https://doi.org/10.2991/978-94-6463-034-3_132

some unreasonable aspects of railway departments can not be adequately solved, which further reduces passengers' ride experience and forms a vicious circle, which is not conducive to the further improvement of the service quality of railway transportation departments.

Therefore, it is of great practical significance to develop a big data analysis system for railway evaluation based on passenger reviews. It is beneficial to the improvement of service quality of the railway transportation department. Based on big data analysis and research, the paper accurately explores the areas that need to be improved in the railway transportation sector, providing a scientific basis and scientific basis for its decision making, effectively avoiding subjectivity, and having positive significance for the railway transportation sector to establish its evaluation feedback system and opinion analysis [3]. This paper conducts a big data analysis and research on passenger-oriented railway service quality based on the LDA model. The contributions of this paper are as follows:

(1) In this paper, the TF-IDF method is used to extract text features. Then, according to the importance of the evaluated words or phrases to the passenger needs and feedback data, the text feature words with high TF-IDF weight are calculated. At the same time, the semantic network is used to analyze the correlation between the main feature words, and according to the central node in the semantic network, the features of crawling data are further found.

(2) This paper uses the LDA topic model for modeling, determines the best topic number of comment data by calculating the confusion degree of the model, and trains the LDA topic model according to the best topic number. Furthermore, the theme frequency index was obtained through the analysis of the model, according to which the corresponding emotional intensity of different themes was obtained. Finally, the effect of the model is verified according to the correlation between different topics.

(3) Finally, according to the results of the LDA theme model, this paper summarizes the existing problems of railway traffic service based on the different demands of passengers and puts forward corresponding improvement suggestions. This paper also uses a neural network to analyze the sentiment of the text data and summarizes the current railway passengers' sentiment tendency.

2 Literature Review

In recent years, scholars at home and abroad have done much work on the analysis and research of text criticism. Therefore, we have carefully read and studied the research done by other scholars related to the topic of this paper and make a summary and review of their work.

Zhang Gongrang et al. proposed in their paper, Text Semantic Mining and Sentiment Analysis Based on Comment Data [4], after data visualization, the correlation among various factors was established through LDA theme model feature analysis, and relevant dictionaries were constructed and reasonably weighted, to get the emotional score of each comment. Different from Zhang Gongrang et al., Jia Ruiyu et al., in their paper, Short Text Clustering Combined with New Concept Decomposition and Frequent Word Sets [5], independently proposed a short text clustering method (CFFIC) combining new concept decomposition and frequent word sets, which realized text clustering faster and better. However, Gu Yongchun et al. in their paper, Quality of Service Evaluation for Unbalanced Network Review Data Mining [6], released that through the methods of subject frequency, subject frequency G index, and heat weighting, the evaluation index system of hotel service quality was constructed, which could effectively alleviate the influence of data imbalance and make the evaluation results more reliable. In terms of the evaluation index system, Ding Yusi et al. in their paper, Research on the Service Quality Evaluation Index System of Five-star Hotels Based on the Content Analysis of Online Review [7], suggested that the evaluation index system was constructed by questionnaire survey and reliability analysis, and the relevant factors determined the index system. Meanwhile, some people also made innovations in the model, and Tao Yongcai et al. independently proposed the MF-MCNN model in their paper, Research on Multi-feature Fusion Method for Short Text Sentiment Analysis [8], which used a multi-channel convolutional neural network to learn and extract more comprehensive emotional semantic information from multi-input features, significantly reducing the training time of the model.

In contrast to the above research work, this paper proposes using the LDA model to conduct big data analysis and research on railway service quality and propose reasonable suggestions for the railway traffic department to serve the passengers better.

3 Method

The LDA topic model is a three-level Bayesian probability graph model consisting of the word-document-topic three-layer structure, as shown in Fig. 1. LDA uses an implicit random variable subject to the Dirichlet distribution to represent the document's subject mix ratio to simulate document generation. One or more topics make up a document, and each word in the document is generated by one of the topics. The model structure is complete and precise, and the dimension represented by the text can be significantly reduced by using the probability inference algorithm to process the text, thus avoiding the dimension disaster. Therefore, as an unsupervised learning technique, LDA is usually used to identify the subject information hidden in a large-scale document collection or corpus. As a result, it has achieved excellent text classification, information retrieval, and other fields.

The specific training process of the LDA model is as follows [9]:

(1) The number of feature words contained N_m in comment m is subject to Poisson distribution, i.e.

$$N \sim P\text{oission}(\xi)$$

The probability density function of *Poisson* distribution is:

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{1}$$

(2) Generate the topic distribution for comments *m*, where m $\in \{1, 2, \cdots, M\}$, θm $\sim$ *Dirichlet*(a) is generated by sampling, where *M* represents the total number of comments in the data set, θ_m represents the topic probability distribution of the m_{th} comment, and α is the prior parameter *Dirichlet* of the multinomial distribution of the topics under each comment.

The probability density function of *Dirichlet* distribution is:

$$f(x1, x2, ..., xk; \alpha 1, \alpha 2, ..., \alpha k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} xi^{\alpha^i - 1} \tag{2}$$

where

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k} \alpha^i)}, \sum xi = 1 \tag{3}$$

(3) Generate the topic $Z_{i,j}$ of the j_{th} word of document *i* from the sample of the polynomial distribution of the topic θ_m.

(4) Sampling from the *Dirichlet* distribution β to generate the word distribution φZi,j corresponding to the topic $Z_{i,j}$, i.e.

$$\varphi\text{Zi,j} \sim D\text{irichlet}\,(\beta)$$

(5) Sampling the polynomial distribution φZi,j of words to generate the selected word subject word item wi, j, i.e. $wi, j \sim Multinomial(\varphi Zi, j)$.

The probability density function of *Multinomid* is:

$$P(x1, x2, ..., xk; n, p1, p2, ..., pk) = \frac{n!}{x1! \cdots xk!} p1^{x1} \cdots pk^{xk} \tag{4}$$

where

$$\sum_{i=1}^{k} pi = 1, pi \geq 0 \tag{5}$$

The generation process of the LDA model is shown in Fig. 1.

In this paper, the Perplexity evaluation index was used to determine the optimal number of topics in the documents. Perplexity is commonly used to measure the degree to which a probability distribution or probability model predicts the merits and demerits of samples and can be used to adjust the number of topics [10]. Its calculation formula is as follows:

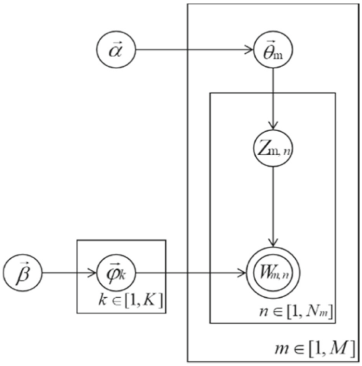

Fig. 1. The generation process of the LDA model

$$Perplexity(D) = \exp\frac{\sum_{d-1}^{M} \log P(Wd)}{\sum_{d-1}^{M} Nd} \tag{6}$$

where D represents the collection of all words in the document; M represents the number of documents; Wd denotes a word in document D; Nd denotes the number of times in each document D; $P(W_{\mathrm{d}})$ indicates the frequency of words in the document. Generally, the *Perplexity* values show a decreasing trend with the increase of potential topics. Therefore, the smaller the *Perplexity* values are, the stronger the generation ability of the topic model is [11]. In conclusion, the topic values with relatively small *Perplexity* and the relatively small number of topics were selected as the optimal model parameters for LDA model training in this paper [12].

4 Example Analysis

4.1 Preprocessing of Comment Statements

This paper uses the LDA model to conduct passenger-oriented big data analysis and research on railway service quality. Sina Weibo and Zhihu, China's mainstream social platform Sina Weibo and the knowledge question-and-answer platform Zhihu are used as data sources. In order to make the obtained data information more effective and targeted, we adopted different processing strategies when using web crawler technology to carry out data crawling on Sina Weibo and Zhihu. As far as Sina Weibo is concerned, after determining a series of keywords and phrases, such as "high-speed railway and suggestion," "high-speed railway and improvement," etc., Python was used to retrieve keywords on the homepage of Sina and crawl the micro-blog contents related to the keywords. Unlike Sina Weibo, because of the unique question-and-answer mechanism of the Zhihu platform, we selected a series of questions related to the theme and then crawled all the answers under the targeted question. Through the data crawling of the two platforms, the primary data sources for analysis and research in this paper were obtained, with 6673 pieces of data.

Table 1. Schematic table of data to be analyzed

Number	***Review***
1	Why can't the KTZ train get into the station with an ID card
2	Face-to-face chairs. You cannot charge your phone. The little table is too small.
3	No seat belts.
4	Separate sleepers for men and women are strongly recommended.
…	…
5454	Suggest a family car for children on the high-speed train. Put all the passengers with children together. I just want to get some sleep.
5455	I sincerely suggest that every bullet train should have a silent compartment.
5456	It is suggested that the high-speed railway should open a special carriage for children, which is convenient for parents and for passengers like me who want to sleep. I have a feeling of trying to sleep many times on May Day.
5457	The salesman keeps selling things on the train. No matter long distance or short distance, there are always carts to sell things, seriously affecting the rest.

In the process of data acquisition, we crawled all relevant data with a web crawler, so there were many noise data in the preliminary data we obtained, such as repeated comments, invalid and irrelevant comments, comments with less information, empty comments, etc., which could not be analyzed and studied directly, so we carried out data cleaning on the preliminary data, including removing comments with repeated, short, blank or a large number of meaningless characters. However, due to the complex semantic expression of Chinese, it was difficult to completely clean data irrelevant to the subject by relying on the computer alone. Therefore, we also adopted manual means to screen and select the crawling data and remove invalid data. After this data cleansing, we obtained a total of 5,457 pieces of data for the final analysis, as shown in Table 1.

4.2 Comment Semantic Mining Based on TF-IDF

Text feature selection is one of the primary problems in the field of text mining and information retrieval. It aims to quantify the feature words extracted from the text and represent the text information with them. In this study, the TF-IDF method was used to extract text features. TF-IDF is a technology used for information retrieval and text analysis. It is usually used to evaluate the importance of a particular word or phrase to a document. A high TF-IDF value indicates that the word has high importance to the document. For example, the higher the frequency of a word in an article, and the lower the frequency of other words in the article. The higher the importance of the word to the current text, the greater the word's TF-IDF value. TF-IDF method is based on such an idea to carry on the realization.

In this paper, the TF-IDF method is used to extract text features from passengers' feedback data after crawling, evaluate the importance of words or phrases to the feedback data of passengers after crawling, and calculate the text feature words with a high weight

Table 2. Top 20 text feature words with TF-IDF values

id	*keywords*	*TF-IDF*	*id*	*keywords*	*TF-IDF*
1	carriage	0.063	11	into	0.030
2	railway	0.057	12	hard seat	0.030
3	destination	0.052	13	transport	0.029
4	tourist	0.052	14	fare	0.028
5	city	0.042	15	sleeper	0.028
6	station	0.040	16	get in	0.027
7	sro tickets	0.039	17	upper berth	0.026
8	seat	0.034	18	fee	0.025
9	child	0.033	19	epidemic	0.024
10	passenger	0.033	20	renaissance	0.024

of TF-IDF. The top 20 text feature values are shown in Table 2. According to the statistical results, the word "carriage" has a high weight of TF-IDF, indicating that it has high importance in passengers' feedback. It is because passengers pay great attention to the experience and feeling in the carriage during railway traffic. That is to say, the carriage service, carriage environment, and carriage experience are the contact points that passengers generally pay more attention to and are easy to produce service complaints, which requires the railway transportation department to pay attention to the optimization of the user experience around the carriage and pay attention to the carriage demands of passengers.

In addition, to visually show the focus and theme of crawling data, this paper draws a word cloud map to realize the visualization of features and generates a word cloud map based on the relatively large number of demands and suggestions proposed by passengers, as shown in Fig. 2.

4.3 Comment Semantic Mining Based on the Semantic Web

Through the analysis of the text feature extraction and word cloud, this article gets some characteristics of the passengers' feedback value. However, given the word, the cloud cannot direct access to the relationship between the main key, so in order to further explore the relationship between characteristic value, also in order to be able to more intuitive understanding of railway transportation departments on the problems existing in the service, In this paper, the TF - IDF to extract the characteristic value of the matrix is built and made a word semantic network analysis, using the generated word matrix and the semantic web, know the correlation between the various characteristic values and logical, excavated the semantic relation between a word, to a certain extent, by the word segmentation caused by the messy text structure relationship integration, In this way, the original text information which a single word item cannot express can be restored. In addition, we can analyze the central nodes in the semantic network to find the features of the comment text.

Fig. 2. Text eigenvalue word cloud map

As can be seen from the co-word matrix in Table 3, time, select, luggage, improve, carriage, chunyun, station, and other words appear more frequently. Meanwhile, the matrix shows the co-occurrence relationship among various high-frequency words. For example, the co-word frequency between “time” and “select” is 587, and the co-word frequency between “time” and “luggage” is 545. As can be seen from the semantic network in Fig. 3, “time,” “select,” and “luggage” have the closest connection with other feature words in the semantic network and the highest frequency of co-words, so they are the three most core feature words in the whole semantic network. This data is also consistent with that in Table 3. In the Semantic Web diagram, besides the words “time,” “select,” and “luggage,” there are several essential nodes: “trouble,” “destination,” “freedom,” and “child.“ The different nodes in the semantic network form an interconnection relationship according to the semantic relationship, which makes the different needs and suggestions of different passengers closely connected so that the railway transportation department can better grasp the needs of passengers and make timely improvements. The results of both figures and tables reveal that the railway transportation sector needs to focus on meeting the needs of passengers in terms of travel schedule, baggage transportation, route selection, and so on.

4.4 Comment Semantic Mining Based on LDA Topic Model

After TF-IDF feature extraction and semantic network analysis, we have been able to find out some factors that affect the clear preference of passengers and the basic relationship between them. However, to further explore the semantics of comment content, we need to use the LDA topic model, a powerful tool for text mining. In this paper, we use the Sklearn library in Python to train the LDA topic model. The integer within the interval [1, 15] is set as the number of candidate topics to obtain the value of the Perplexity of different models, as shown in Fig. 4.

Figure 4 shows the degree to which the document is uncertain about each potential topic. The lower the degree of confusion, the stronger the generating ability of the topic model. That is, the better the clustering effect of the model. Generally speaking, with the

Table 3. Part Of Text eigenvalue coword matrix

	time	***select***	***luggage***	***improve***	***carriage***	***chunyun***	***station***
time		587	545	-	218	176	257
select	587	-	321	-	-	-	-
luggage	545	321	-	-	-	-	-
improve	-	-	-	-	-	210	-
carriage	218	-	-	-	-	-	-
chunyun	176	-	-	210	-	-	-
station	257	-	-	-	-	-	-

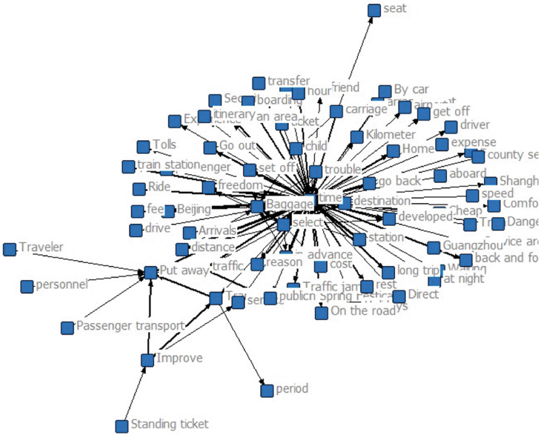

Fig. 3. Passenger feedback eigenvalue semantic network graph

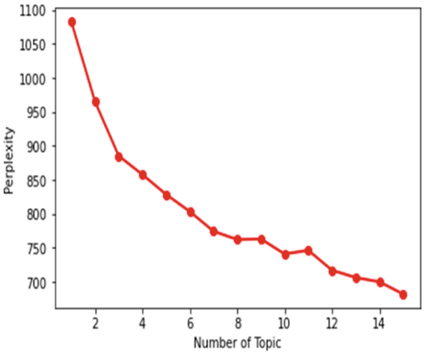

Fig. 4. Perplexity - topic line chart

Table 4. LDA topics and the distribution of thematic keywords

id	*feature words*				
topic1	time	luggage	select	unnecessary	destination
topic2	carriage	child	second-class	passenger	seat
topic3	improve	tourist	service	chunyun	pack-up

increase of the number of potential topics, the Perplexity value will decrease fluctuation. In the model in this paper, the local maximum point of Perplexity appears on the model selection with a topic number of 9. In the original downward trend, when the number of topics changes from 8 to 9, the Perplexity value does not decrease but increases. The more topics, the more complex the subsequent topic analysis will be. Therefore, according to Occam's razor principle, eight potential subject numbers are selected in this paper.

After determining the optimal number of topics, this paper, based on the Python language machine learning package Sklearn, conducts LDA topic modeling on the comment data and gets eight topics and their topic feature word distribution. In the future, we will also analyze the relationship among the eight topics. However, to show the modeling effect, only the first 3 topics are shown here, and the distribution of the first 5 characteristic vocabularies of each topic is shown in Table 4.

LDA topic mining can be divided according to semantics, and several implicit topics expressed by semantically related words can be obtained. For example, Topic 1's vocabulary set describes the topic "accessibility and convenience," Topic 2's vocabulary set describes the topic "ride experience," and Topic 3's vocabulary set describes the topic "vehicles and ticketing services."

Now that we have a trained LDA model, we do not know how the topics relate to each other and how vital each potential topic is in all the online comments. To do this, we will use a Python visualization package called PylDavis to understand the relationships between topics better. The PylDavis package is designed to help users interpret topics that fit into text datasets in a topic model. It can be used to Plot an Intertopic Distance Plot to help users understand the relationships between topics, including the underlying high-level structures between topic groups [13].

The PylDavis package calculates the frequency of each topic, denoted by the size of the circle, and numbered sequentially from 1 to n. So the size and number of the bubbles indicate the frequency of the topic.

The subject spacing diagram drawn is shown in Fig. 5.

According to the above figure, we rearrange the original 8 topics and their distribution of feature words according to the occurrence frequency of the topics and get the topic-feature words table, as shown in Table 5.

According to the results, we calculated the user comment topic strength graph, as shown in Fig. 6. Among them, topic 1 accounted for as high as 29%, while topic 8 only accounted for about 6%.

The analysis shows that the "accessibility and convenience" expressed in theme 1 occupy the primary position in the theme intensity of passenger feedback, indicating that

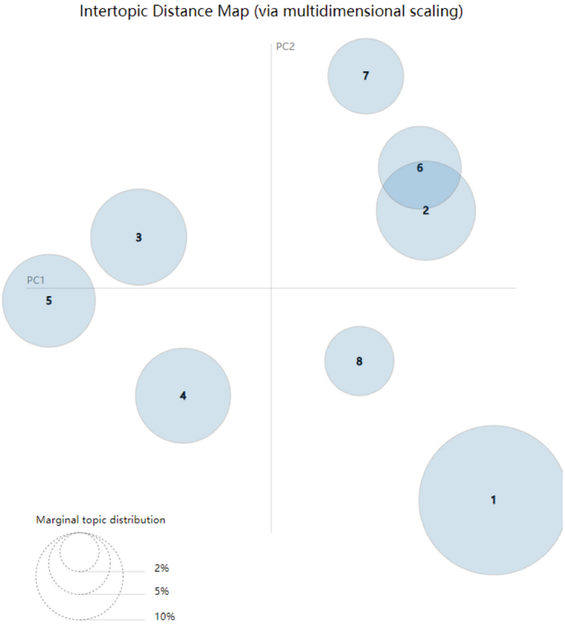

Fig. 5. Intertropical Distance Map

Table 5. The topic-feature words table

topic1		**topic2**		**…**	**topic8**	
key words	***weight***	***key words***	***weight***	**…**	***key words***	***weight***
time	0.233	carriage	0.155	…	time	0.037
luggage	0.018	child	0.064	…	home	0.03
select	0.016	second-class	0.036	…	need	0.011
…	…	…	…	…	…	…
out	0.003	position	0.004	…	special	0.004
cost	0.003	night	0.004	…	reason	0.004
county	0.003	feeling	0.004	…	night	0.004

it plays a relatively important role in the railway transportation experience and demand of passengers, and it is also an urgent need for railway transportation departments to improve the quality of railway service. As far as "accessibility and convenience" is concerned, which is the most urgent theme of passenger demand, railway departments can implement improvement measures from these aspects according to the theme feature words, such as "baggage transportation," "travel route," and "travel cost." In this way, the research on measures based on the theme and its feature words can make the service work of the railway more targeted and effective.

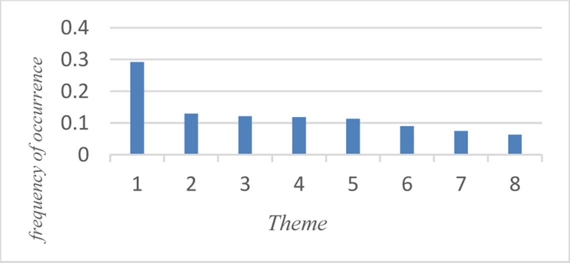

Fig. 6. Passenger feedback theme strength chart

At the same time, the PylDavis package uses multi-dimensional scaling analysis to extract the principal components as dimensions and distributes the topics on these two dimensions. The distance between the topics expresses the proximity between the topics. The bubble distance is JSD distance, which can be considered as the degree of difference between topics. The overlap of bubbles indicates that the feature words in the two topics cross. As shown in Fig. 1, the feature words for topic 2 and topic 6 are partially overlapped, and the information contained in the two topics is likely similar. Through analysis, the feature words of both subjects contain “carriage,” “seat,” and “hard seat,” which indicates that the feature words that appear in multiple subjects such as “carriage,” “seat,” and “hard seat” have influences on multiple subjects. Then these cross keywords refer to the content of railway service, which the railway transportation department should focus on, such as “carriage environment,” “seat comfort,” and so on.

However, the JSD distance between other topics is relatively far, and each topic is independent of each other, which also reflects that the LDA model trained in this paper has a better effect of classifying and summarizing text semantics and can mine a large number of mutually independent semantic potential topics in-text comments.

4.5 Sentiment Analysis

In this paper, when analyzing the current railway passengers’ emotional tendency of online reviews, Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis (SKEP) is used to train Bi-directional Long Short-Term Memory (BILSTM) neural network, written in Python. Skep is an emotional pre-training algorithm based on emotional knowledge enhancement proposed by the Baidu research team. This algorithm uses an unsupervised method to mine emotional knowledge automatically and then uses emotional knowledge to build pre-training targets so that machines can learn to understand emotional semantics [14]. BILSTM neural network trained in SKEP method is used in this paper to obtain sentiment analysis results, as shown in Table 6.

According to the positive and negative probability of emotion calculated by the model, we screened out the comments with the positive probability of emotionless than 0.005 and finally obtained 1,185 relevant comments as the new data set. This paper believes that this part of the review data contains the most negative feelings of passengers and is related to passengers’ issues most eager to solve. In this paper, the LDA model was used to analyze this part of data further, and it was found that the target

Table 6. Proportion of each affective tendency

emotional tendency	***Statistics on Comments***	***proportion***
positive	1134	20.78%
negative	4323	79.22%
total	5457	100%

demands of passengers focused on the following two topics, "pre-ride experience" and "ride experience." The former focuses on passengers' difficulty in buying tickets and inconvenient luggage transportation, while the latter shows passengers' dissatisfaction with seats and noisy children in trains. It requires the railway authorities to increase the frequency of trains on important routes to relieve the pressure on passengers to buy tickets and improve luggage transportation. On the other hand, they have to come up with solutions for the comfort of seats in carriages and the noisy children. Based on the analysis of strong negative emotions, the transportation department can better understand the urgent needs of passengers, make adjustments to urgent problems, and then optimize the service quality.

According to the above characteristic analysis results, the main factors that domestic passengers are not satisfied with railway transportation are accessibility and timeliness, service attitude, ticketing business, peak transportation, facility construction, epidemic prevention and control, complaint handling, and competition with other modes of travel. Therefore, we put forward the following suggestions:

(1) Further improve the speed of trains and reduce the probability of delayed trains, and build more stations and transport routes on the original basis. At the same time, railway departments should be fully prepared to respond to the rush period transport stress measures to ensure that people's essential requirements for travel during holidays are met.
(2) Pay more attention to service quality and improve the quality of train crew. In selecting and employing persons, railway departments should assess and cultivate comprehensive quality and service consciousness and establish adequate real-time supervision of service quality. In addition, to ensure that passengers remain relaxed and happy during the whole ride.
(3) The internal facilities of the train should be upgraded, unreasonable design parts should be deleted, and the design of facilities such as seats, charging ports, and luggage racks should be improved to ensure that these facilities can serve passengers well and give passengers a good ride experience.
(4) Optimize the ticketing system to provide more diversified ticketing methods so that passengers can experience the same convenience when purchasing tickets and ensure that passengers will not encounter problems such as finding a ticket.

Through big data analysis and prediction, and understanding of passengers' needs and expectations, railway services can be improved in a targeted way. Therefore, the railway department needs to analyze and understand passengers' demands through big data to optimize the service process further and improve service quality.

5 Conclusion

This study takes comment text and data from Sina Weibo and Zhihu as examples and uses word cloud map and semantic network feature correlation analysis to conduct feature analysis on passenger comment text data of high-speed railway. This paper uses the evaluation index of the degree of confusion to determine the optimal number of topics in the LDA model, constructs the LDA topic model, calculates the topic frequency, and explores the relevance between topics. In order to explore passengers' emotional tendencies, this paper constructs a bi-directional long and short-term memory neural network (BILSTM) to conduct dynamic analysis on online reviews. It then selects the comments with the deepest negative emotions for semantic mining to find out the two aspects of high-speed rail issues that passengers are most concerned about at present. This study focuses on the most popular issues and analyzes the current situation of the railway transportation industry based on user comments. At the same time, on the basis of the research results, five suggestions are put forward, which are of great significance to improve the service quality of railway departments, meet the expectations of passengers, and improve the overall service level of railway transportation industry.

In addition, this paper has some limitations in the study: This article only according to the "high-speed rail," "advised" the topic of the two typical data collection and analysis, in the follow-up study, will further expand the railway transportation industry data collection and analysis of the other topics, many under the topic of comparison research, so this paper adopts a series of analysis method has better universality, To the adjustment of the railway sector, improve to have a better guiding effect.

References

1. Yu Chengcheng, Lv Hongxia. Research on Passenger Transport Marketing Strategy Based on Big Data Railway Passenger Portrait [J]. Journal of the China Railway Society,2020,42(08):23-28.
2. Gan Ziqin. Research on the Service Quality of Tourism E-Commerce Based on Online Review Mining -- Taking Ctrip as an Example [J]. China's Collective Economy,2021(13):71-73.
3. ZHANG Junfeng. Design and Application of Railway Passenger User Portrait System [J]. Railway Computer Application,2018,27(07):54–57.
4. Zhang Gongrang, Bao Chao, Wang Xiaoyu, Gu Dongxiao, Yang Xuejie, Li Kang. Text Semantic Mining and Sentiment Analysis Based on Comment Data [J]. Information Science,2021,39(05):53-61.
5. Jia Ruiyu, Chen Shengfa. Short Text Clustering Combined with New Concept Decomposition and Frequent Word Sets [J]. Journal of Small and Microcomputer Systems,2020,41(06):1321–1326.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
6. Gu Yongchun, Gu Xingquan, Wu Jiao, Hong Caifeng, Jin Shiju. Quality of Service Evaluation for Unbalanced Network Review Data Mining [J]. Small and Micro Computer System, 201,42(02):354–361.minutes.

7. Ding Yusi, Xiao Yinan. Research on the Service Quality Evaluation Index System of Five-star Hotels Based on the Content Analysis of Online Review [J]. Consumer Economics,2014,30(03):64–69.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
8. Tao Yongcai, Zhang Xinqian, Shi Lei, Wei Lin. Research on Multi-feature Fusion Method for Short Text Sentiment Analysis [J]. Journal of Small and Micro Computer Systems,2020,41(06):1126-1132.
9. ZU Xian. Review of the latent dirichlet allocation topic model[J]. Journal of Hefei Normal University, 2015,33(6): 55-58.
10. Hagen L. Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models? [J]. Information Processing & Management, 2018, 54(6): 1292-1307.
11. Du Y J, Yi Y T, Li X Y, et al. Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation[J]. Engineering Applications of Artificial Intelligence, 2020, 87: 103279.
12. Ma T H, Li J, Liang X N, et al. A time-series based aggregation scheme for topic detection in Weibo short texts[J]. Physica A: Statistical Mechanics and Its Applications, 2019, 536: 120972.
13. Sievert, Carson & Shirley, Kenneth. (2014). LDAvis: A method for visualizing and interpreting topics. https://doi.org/10.13140/2.1.1394.3043.
14. Tian H , Gao C , Xiao X , et al. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis[J]. 2020.