Hindawi
Complexity
Volume 2021, Article ID 9126745, 13 pages
https://doi.org/10.1155/2021/9126745

Research Article

Fault Diagnosis of Signal Equipment on the Lanzhou-Xinjiang High-Speed Railway Using Machine Learning for Natural Language Processing

Lei Shi,[1] **Yulin Zhu**,[1] **Youpeng Zhang**,[1,2] **and Zhongji Su**[1]

[1]*School of Automation & Electrical Engineering, Lanzhou Jiaotong University, Lanzhou, China*
[2]*Rail Transit Electrical Automation Engineering Laboratory of Gansu, Lanzhou Jiaotong University, Lanzhou, China*

Correspondence should be addressed to Yulin Zhu; zhuyl1228@foxmail.com

Received 29 June 2021; Revised 15 July 2021; Accepted 20 July 2021; Published 28 July 2021

Academic Editor: Muhammad Javaid

Copyright © 2021 Lei Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Lanzhou-Xinjiang (Lan-Xin) high-speed railway is one of the principal sections of the railway network in western China, and signal equipment is of great importance in ensuring the safe and efficient operation of the high-speed railway. Over a long period, in the railway operation and maintenance process, the railway signaling and communications department has recorded a large amount of unstructured text information about equipment faults in the form of natural language. However, due to irregularities in the recording methods of these data, it is difficult to use directly. In this paper, a method based on natural language processing (NLP) was adopted to analyze and classify this information. First, the Latent Dirichlet Allocation (LDA) topic model was used to extract the semantic features of the text, which were then expressed in the corresponding topic feature space. Next, the Support Vector Machine (SVM) algorithm was used to construct a signal equipment fault diagnostic model that reduced the impact of sample data imbalance on the classification accuracy. This was compared and analyzed with the traditional Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbor (KNN) algorithms. This study used signal equipment failure text data from the Lan-Xin high-speed railway to conduct experimental analysis and verify the effectiveness of the proposed method. Experiments showed that the accuracy of the SVM classification algorithm could reach 0.84 after being combined with the LDA topic model, which verifies that the natural language processing method can effectively realize the fault diagnosis of signal equipment and has certain guiding significance for the maintenance of field signal equipment.

1. Introduction

Railway signal equipment mainly includes railway signals, station interlocking equipment, and section blocking equipment. The main function of these types of equipment is to ensure the safety of train operation and shunting work and increase the capacity of the railway. At the same time, it also plays an important role in increasing the economic benefits of railway transportation and improving the labor conditions of railway workers. With the rapid development of information technology and the upgrading of railway signal equipment, the railway bureau has accumulated a large amount of signal equipment failure data. However, the descriptions of the equipment faults in these failure data are mostly recorded in unstructured natural language. Therefore, in the process of dealing with equipment failure, maintenance personnel still use a combination of expert knowledge and personal experience to diagnose the problems. This method usually brings great challenges to on-site technicians to accurately diagnose the faults due to a possible lack of experience of the personnel and an incomplete understanding of the on-site conditions. At the same time, it also causes significant safety hazards due to delays in handling the faults. Therefore, in the era of big data, the application of machine learning and natural language processing methods to diagnose railway signal equipment faults can reduce the technical demands on field maintenance staff. It is of great significance to improve the efficiency of railway signal equipment fault diagnosis and ensure the safe and efficient operation of the railways.

At present, the fault records of railway signal equipment in China consist of unstructured Chinese short text recorded in natural language. They contain a large number of specialized professional vocabulary mixed with numbers, letters, and some special symbols, making it difficult to perform word segmentation and feature extraction. Until now, there has been little research on the fault text mining of railway signal equipment both at home and abroad. Zhao and Xun [1] used the method of text mining to diagnose the faults of on-board equipment and used the traditional Bayesian network method as the fault classification method. Yang et al. [2] used the Synthetic Minority Oversampling Technique (SMOTE) algorithm to intelligently classify unbalanced text data, which solved the problem of sample imbalance to a certain extent. Zhong, Tang, and Wang[3] studied the feature extraction and diagnosis of turnout faults, and ShangGuan et al. [4] used the topic model method to perform fault diagnosis on vehicle equipment. Because the equipment fault recorded in the fault record table is not consistent between each station or section of railway, it has a certain impact on the number and category of post-processing and statistical faults. Some scholars have proposed [5, 6] to solve the problem of inconsistent maintenance record standards for on-site personnel through standardized documents, which to a certain extent promotes the standardized management of railway big data records. However, there have been no studies on how to effectively interpret such a large amount of analyzed and processed data.

In this paper, the topic model method was used to extract the semantic features of the fault text recorded by the railway signaling and communications department. The topic model is a kind of statistical model that clusters the implied semantic structure of texts set by unsupervised learning. In reference [7], the Latent Dirichlet Allocation (LDA) model was used to mine topics of conference summary text, with different topics being clustered. Peng et al. [8] used the LDA topic model to extract product features in social media reviews and studied the relationship between different reviews and customer sentiment. References [9–11] introduced the method of the topic model to deal with the task of text classification, and all achieved satisfactory results. Additionally, the topic model can extract the semantic information of text effectively and find the potential correlation information between each document and vocabulary in the text [12].

Based on the analysis of the features of the fault record text, this study used an algorithm based on machine learning and natural language processing and the method of the LDA topic model to extract the word item features and the theme features of the corresponding fault in the fault description text for railway signal equipment. It transformed the corresponding fault document into a theme feature space model, which reduced the dimension of the feature effectively and made the reduced dimension data easier to process and use. Because the recorded fault data had the characteristics of unbalanced distribution, the Support Vector Machine (SVM) classifier was selected to classify the faults. The SVM classifier is not sensitive to unbalanced distribution data and is recognized as one of the most effective models for processing small data samples.

2. Fault Text Analysis of Railway Signal Equipment

Railway signal equipment mainly includes dispatching centralized traffic control (CTC) system equipment, train dispatching command system (TDCS) equipment, train control equipment, interlocking equipment, switch equipment, track circuits, signal machines, snow melting equipment, power supply equipment, etc. There are also many kinds of classification methods for railway signal equipment faults. In this paper, according to the function of the corresponding equipment and the mode of fault phenomenon, it is divided into 10 kinds of faults. These are switch equipment fault, track circuit fault, signal machine fault, snow melting equipment fault, train control equipment fault, CTC equipment fault, TDCS equipment fault, microcomputer interlocking fault, power supply equipment fault, and other faults. To verify the correctness and effectiveness of the model, 1000 fault text data samples recorded by the Lan-Xin high-speed railway signaling and communications department were selected for experimental analysis. Figure 1 shows the distribution of signal equipment failures recorded by the Lan-Xin high-speed railway signaling and communications department from 2017 to 2019.

2.1. Fault Record Data of Signal Equipment. In this study, the fault record data provided by the Lan-Xin high-speed railway signaling and communications department were screened manually. Some fault phenomena that were not recorded in detail, fault diagnosis results that were recorded incorrectly, or unrecorded fault data were all manually removed, making the selected data more conducive to follow-up feature extraction and classification work. Table 1 shows some examples of fault records.

It can be seen from Table 1 that there is a large amount of fault feature information in both the fault overview and cause analysis columns. However, due to differences in personal language habits and maintenance experience of field maintenance personnel, the data forms recorded in natural language have certain differences, which makes the subsequent fault feature extraction and fault diagnosis algorithm more complex and difficult.

2.2. Signal Equipment Fault Diagnosis Process. After the fault is discovered, the on-site maintenance personnel usually take photos of the location where the fault is found and make a detailed text record of the situation when the fault occurs, and then use a combination of personal experience and expert knowledge to diagnose the type of the fault. Figure 2 is a photo of part of the fault location on-site.

Figure 3 shows the process of signal equipment fault diagnosis, which mainly includes fault word item feature extraction, fault topic feature extraction (implied semantic feature), and fault diagnosis. Because Chinese text is different from English text, in which words are separated by

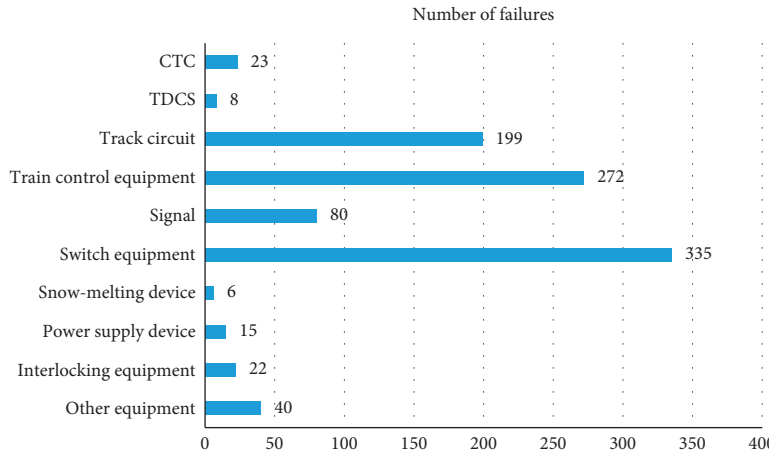

Figure 1: The distribution of signal equipment failures recorded by the Lan-Xin High-speed Railway Signaling and Communications Department from 2017 to 2019.

spaces, it is necessary to deal with text word segmentation first and then extract the word item features using the bag-of-words (BoW) model. Because of the high dimension of fault records and the loss of semantic information caused by the great difference in the details of recorded fault data, great challenges were encountered in the fault diagnosis. Therefore, this study adopted the topic model method to extract fault features and reduce the dimension of the data and express the fault document in the topic feature space. Finally, the support vector machine was used as the fault diagnosis device to diagnose the signal equipment fault.

3. Fault Text Segmentation and Fault Feature Extraction of Railway Signal Equipment

Proper data preprocessing can effectively reduce the complexity of the algorithm for signal equipment fault diagnosis and improve the accuracy of the diagnosis. In the field of natural language processing, the traditional method usually expresses the original text data in the vector space model (VSM) of word items. The dimension of the vector is the number of words. This model ignores word order, grammar, and syntax and loses part of the semantic meaning. It only counts the number of times each term appears in different documents. For the fault text of railway signal equipment in this paper, its original feature is the fault word feature. However, using the bag-of-words model to express the signal equipment fault will ignore the relationship between the feature of fault words and the topic, so it will reduce the accuracy of the subsequent fault diagnosis.

In recent years, a statistical model for discovering hidden topics in a series of documents has been widely used in many fields, such as user evaluation, social media, emotional classification, medical evaluation, user preferences, etc. [13–17]. Therefore, to improve the shortcomings of the traditional bag-of-words model and improve the automatic degree of feature extraction and adaptability for later fault diagnosis, this study used a combination of Chinese text word segmentation [18, 19] and topic models [20, 21] to extract the theme after word segmentation. The steps of word segmentation and feature extraction are shown in Figure 4.

3.1. Generation of Fault Lexicon. Since the original fault text of signal equipment is recorded in the form of natural language, it is necessary to use Chinese word segmentation technology to process the original document. The core process of the Chinese word segmentation algorithm is to input the Chinese text to be segmented, then identify the stop vocabulary to remove the useless words in the document, use the user-defined dictionary to correctly segment specific vocabulary, and finally output the optimal word segmentation result. The flow chart of the Chinese word segmentation algorithm is shown in Figure 5.

In this study, the Jieba text segmentation package composed in the Python language was used to segment the signal fault text. The Jieba algorithm uses a prefix dictionary to achieve efficient word graph scanning and generates a directed acyclic graph (DAG) composed of all possible generations of Chinese characters in a sentence. Then it uses dynamic programming to find the maximum probability path, and based on word frequency, finds the largest segmentation combination. For unknown words, a Hidden Markov Model (HMM) based on Chinese character word-forming ability was adopted, and the Viterbi algorithm was used. The general vocabulary of the Chinese word segmentation tool does not contain a professional vocabulary in the railway signal field. Therefore, if it was directly used to segment the data in this paper, it would lead to incorrect results or undivided words, which would make the result of segmentation unsatisfactory, and then affect the subsequent feature extraction.

As a result, the lexicon related to railway signal equipment faults needed to be manually selected. In principle, all the word items in the document could be used for training, but some of them had no essential connection with the fault. For example, some station names, staff member names, "and," "or," and other words and punctuation marks needed to be de disabled. At the same time, some words in the field of railway signals, such as a red-light strap, point machine,

Table 1: Example of signal equipment fault data.

Num	Railway section	Department	Fault location	Equipment type	Failure overview	Cause analysis
1	Lanzhou	Minhe South	Qin'an to Tongwei	Switch equipment	At 21:25 on March 14, 2019, 7 turnouts positioning of Minhe south station was not indicated, which affected the d2708 stops of Minhe south station for 10 minutes. After on-site treatment, 21:51 positioning indicated restoration and 21:56 distribution record.	During the centralized maintenance of Minhe South station on March 14, on-site operators failed to grasp the seasonal changes and improperly adjusted the gap of 7#*X*1, and the card gap caused no indication of the positioning of the 7# turnout.
2	Jiayuguan	Jiayuguan South	Jiuquan South	Switch equipment	At 16:28 on April 6, 2019, no switch reverse indication in no. 6 turnout of Jiuquan South station of Lanzhou Xinjiang passenger dedicated line. After the on-site inspection and treatment by the staff, the equipment was opened and returned to normal use at 17:19.	When the no. 6 turnout J1 machine is switched from the positioning to the reverse position, the lock hook is not strong, resulting in no switch reverse indication of the no. 6 turnout reverse position.
3	Lanzhou	Lanzhou West	Lanzhou West railway station	Train-control equipment	At 9:38 on April 21, 2019, the CRH380B-5734 EMU of the Lanzhou Depot of the Lanzhou bureau Group Company was responsible for the 0G846 train traction task. When leaving the station, the CRH380B-5734 EMU was caused by "the emergency braking of whole trains was not relieved" the group cannot leave.Replaced with CRH380B-5641 EMU to take on the 0G846 train traction task. CRH380B-5641 EMU departs at 10:16, leaving 38 minutes late.	The code sending signal of the locomotive signal code box is abnormal, causing the EMU to display an emergency braking state.
4	Jiayuguan	Yumen	Liugou South	Switch equipment	At 5:14 on June 16, 2019, no. 6/8 turnout of Liugou South station was reversed and there was no switch reverse indication. After inspection and test, it was opened at 6:07 and the equipment was restored to normal use.	The lock hook on the reverse side of the no. 6 *X*2 machine is not strong, causing no switch reverse indication of the reverse position of the no. 6/8 turnout.
5	Lanzhou	Lanzhou West	High-speed trains yard	Track circuit	At 22:36 on August 21, 2019, when the *D*1 track car was driven into the high-speed field from the EMU, the 401DG and 235DG lost train occupation prompt.	On-site inspection at 23:10 revealed that the rail surfaces of the two track sections of 235DG and 401DG were rusted severely, the rail wheel pairs were rusted, and the rail car was lighter and failed to fully fit the rail surface. These three factors led to the track circuit. The shunt is poor, causing the loss of the *D*1 track car occupation prompt.

manual release, block, and other words that have important relationships with fault types, did not exist in the dictionary of the general domain. In the process of word segmentation, it was necessary to establish a user-defined vocabulary list in the field of railway signaling according to expert knowledge. The corresponding words were added to the vocabulary list so that the segmentation results could be more accurate. Figure 6 is a sample of some vocabulary in the field of railway signaling. After analysis and summary, a total of 98 related terms were summarized.

After the Chinese word segmentation was performed on the signal fault documents, the corresponding railway signal fault dictionary could be obtained. Afterward, the unstructured text was represented as a structured vector through the VSM model. The generation of the term-document matrix is shown in Figure 7.

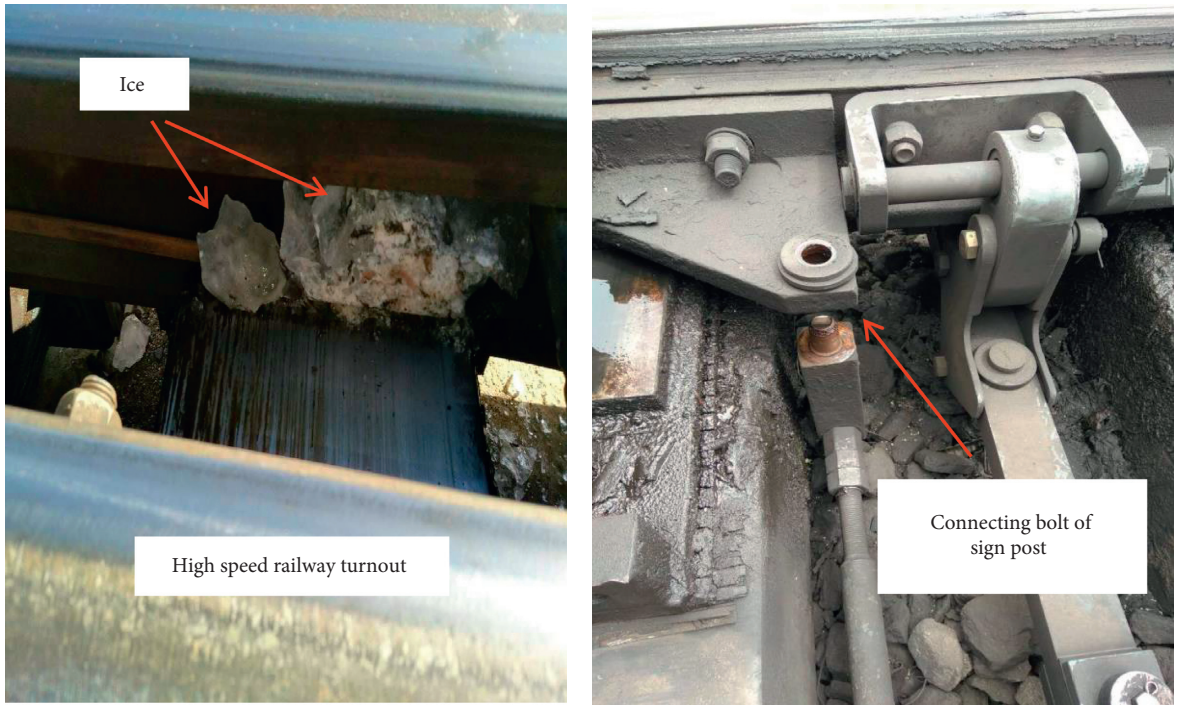

Figure 2: (a) Icing between the turnouts; (b) The connecting bolt of sign post's joint is broken.

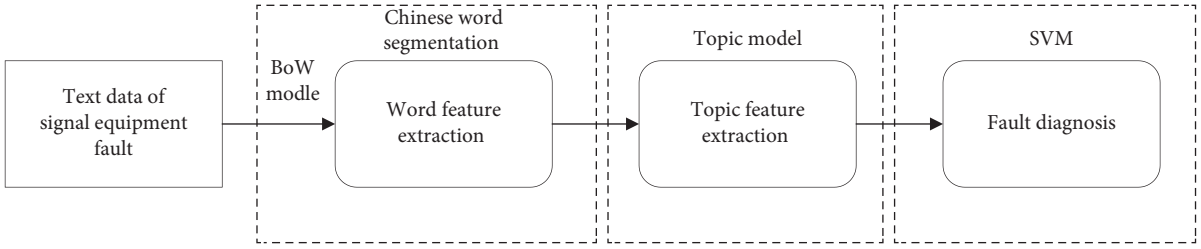

Figure 3: Signal equipment fault diagnosis process.

3.2. Fault Feature Extraction of Railway Signal Equipment Based on the LDA Topic Model. Because the traditional word feature space dimension will increase with an increase in the word list size, the vector is quite sparse. Secondly, the word feature space cannot deal with the problems of "one-word polysemy" and "one meaning many words," which greatly increases the complexity of subsequent fault diagnosis. To combat this issue, the topic model method has developed rapidly in recent years [22–24]. The topic model tries to realize the representation of the text from the perspective of the probabilistic generative model. Each dimension is a "topic," and this topic is usually a cluster of words. Therefore, it is possible to guess the semantics represented by each dimension through the topic. It is explanatory and can transform the document from the lexical feature space to the topic feature space so that the shortcomings of the traditional lexical feature space can be solved. This study used the LDA topic model algorithm to extract the features of railway signal equipment fault records.

3.2.1. LDA Model. The LDA topic model is a Bayesian model of the Probabilistic Latent Semantic Analysis (PLSA) algorithm. It is composed of a three-layer structure of documents, topics, and words. It is an unsupervised machine learning technology and is often used in text topic recognition, text classification, and text similarity calculation. It assumes that a document has multiple topics, and each topic corresponds to different words. The process of constructing a document is to first select a topic with a certain probability and then select a word under this topic with a certain probability, so that the first word of the document is generated. If this process is repeated continuously, the whole

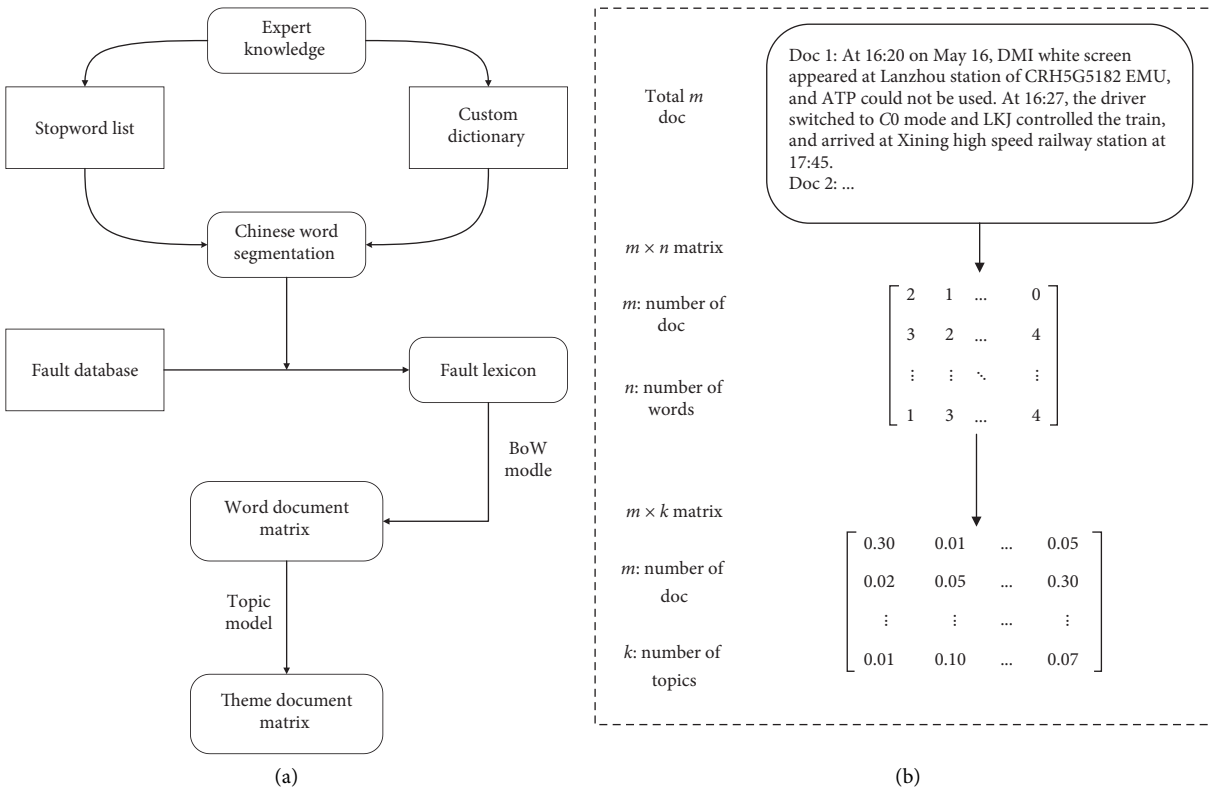

Figure 4: (a) Fault text segmentation and feature extraction process; (b) Feature extraction process based on topic model.

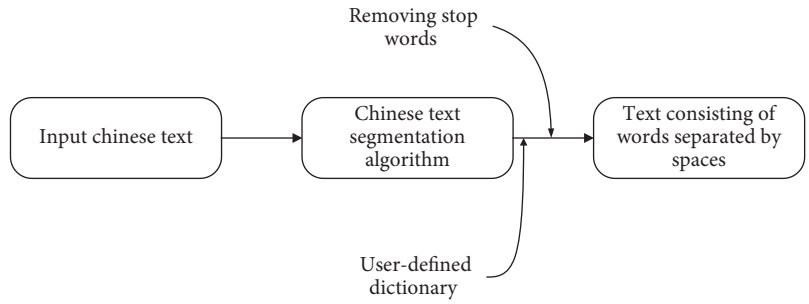

Figure 5: Chinese word segmentation algorithm flow.

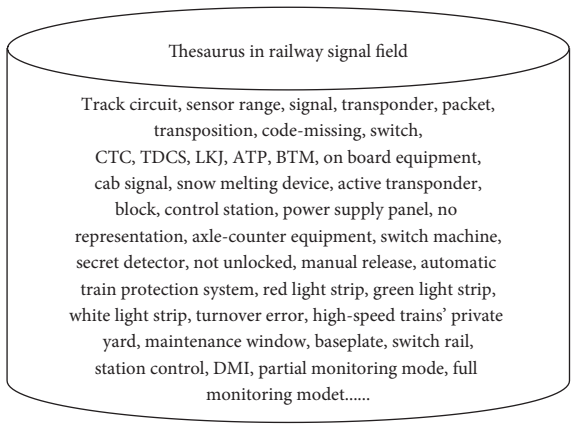

Figure 6: Railway signal domain Thesaurus (part).

article will be generated. From documents to topics, and topics to terms, all obey the polynomial distribution. The probability graph model of the LDA model is shown in Figure 8.

In Figure 8, M represents the total number of documents included in the training, and each document has N number of words, K represents the number of topics, θ_i is the main body distribution of document d_i, and $\varphi_{z_{i,j}}$ is the word distribution of topic $z_{i,j}$. Dirichlet (α) is the prior distribution of parameter α, which is used by the LDA model to sample and generate the topic distribution corresponding to the document. Dirichlet (β) is the prior distribution of parameter β, and the word distribution corresponding to the topic is generated by sampling in the model [25]. The relationships of all variables in the model are as follows:

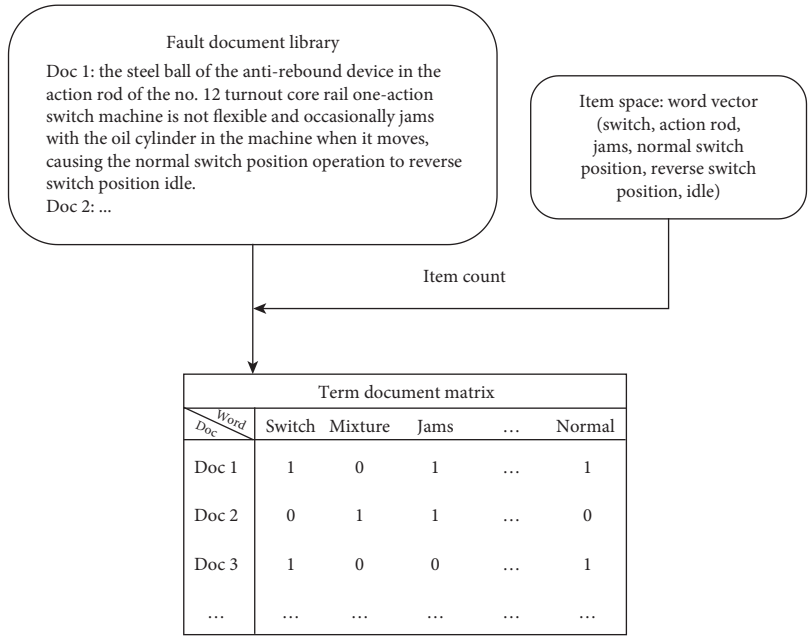

Figure 7: Expression process of fault document in term space.

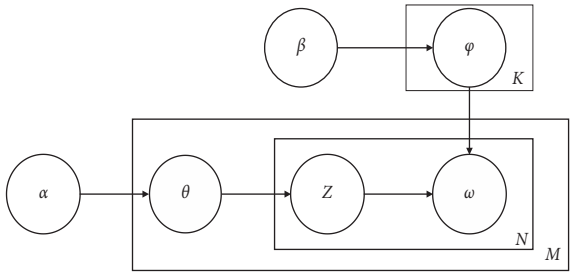

Figure 8: LDA probability graph model.

$$p(w_i, z_i, \theta_i, \varphi|\alpha, \beta) = \prod_{i=1}^{N} p\left(w_{i,j}|\varphi_{z_{i,j}}\right) p\left(z_{i,j}|\theta_i\right) \cdot p(\theta_i|\alpha) p(\varphi|\beta). \tag{1}$$

By using the joint probability distribution, the conditional distribution of hidden variables under a given observation variable value is calculated [26].

$$p(w_i|\alpha, \beta) = \int_{\theta_i} \int_{\varphi} \sum_{z_i} p(w_i, z_i, \theta_i, \varphi|\alpha, \beta). \tag{2}$$

When analyzing the original data, the entire document set is used as the input content for LDA training to obtain the topic distribution of each record.

3.2.2. Feature Extraction of Railway Signal Equipment Fault Subject. When the topic model is applied to extract the theme features of railway signal equipment faults and transform them into the topic document matrix, the number of subject features K must be input. Generally, the size of the K value needs to be given before training the corresponding model, and there is no fixed method to determine it. Therefore, a certain degree of subjective experience is needed to determine the value of topic number K. The simple method to determine the optimal K value is usually repeated experiments with different K values. When the evaluation function (such as the precision of the classifier) reaches the optimal value, the K value is considered to be optimal. If the value of K is too small, the precision of classification will be relatively low, and when the value of K is too large, the purpose of denoising will not be achieved. Figure 9 shows the change of SVM classifier precision for different numbers of topics. In 20 experiments, the precision rate twice exceeded 80%: 0.801 at $K = 10$ and 0.840 at $K = 17$. Therefore, $K = 17$ was selected as the optimal number of topics.

After the value of the number of topics K was determined, the LDA topic model was used to perform dimensionality reduction and feature extraction processing on the term-document matrix. After dimensionality reduction of the term-document matrix, the components in the topic space had semantic meaning and corresponding characteristics. Table 2 shows the results of fault records for signal equipment of the Lanzhou-Xinjiang high-speed railway after feature extraction using the LDA topic model.

It can be seen from Table 2 that Topic $T1$ is related to the fault of "loss of indication of a switch," Topic $T2$ represents a "switch idling" fault, Topic $T3$ is related to a "transponder fault," and Topic $T4$ corresponds to a fault in "wireless communication." Figure 10 illustrates the expression process of the fault document in the topic feature space.

4. Fault Diagnosis of Railway Signal Equipment Using the Support Vector Machine

The support vector machine (SVM) is a widely used classifier method in classification and regression analysis and has achieved satisfactory results in many applications [27–29].

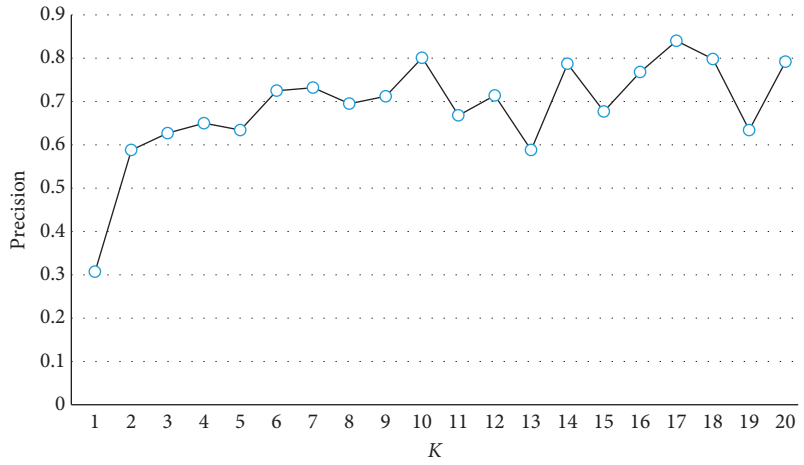

Figure 9: Changes in classification precision under different K values.

Table 2: Example of signal equipment fault data.

Topic	Terms included
$T1$	Switch, normal position, reverse position, loss of indication
$T2$	Idle, iron filings, baseplate, switch rail, scraping
$T3$	ATP, transponder, deletion, abnormal data, information loss
$T4$	Communication, $C2$, $C3$, wireless connection, timed out, ATO
$T5$	Restart, DMI, black screen, change, auxiliary screen, white screen
$T6$	Overrunning, emergency braking, linkage, partial monitoring
$T7$	Loss of occupancy, CTC, flash off, crash, trigger
$T8$	Dispatch, loss of occupancy, wrong train number
$T9$	Red light strip, materials, bad insulation, fuse
$T10$	Mixed line, internal module, I-V curves, red light strip
$T11$	Switching performance, trip, ground wire, circuit breaker, current limiting
$T12$	Slot insulation, iron filings, voltage levels drop, short-circuit, sundries
$T13$	PIO, machine cabinet, terminal, control station, black screen
$T14$	Lights out, lighting unit, contact problems, broken wire
$T15$	DMI, communication interruption, black screen, train-control, start up
$T16$	Snowmelt, heating, voltage, snowfall, temperature
$T17$	UPS, power supply, cut-out, high temperature, battery

There are many kinds of faults in railway signal equipment, the distribution of the various faults is unbalanced, and the sample number of some faults is very limited. Support vector machines are advantageous when dealing with small sample data sizes and are less sensitive to unbalanced data. The SVM can rely on a small number of support vectors to perform the corresponding diagnosis. Therefore, this study used the learning algorithm of the support vector machine to diagnose the faults of railway signal equipment.

4.1. Description of the SVM Algorithm. The SVM method is based on the Vapnik-Chervonenkis (VC) dimension theory of statistical learning theory and the principle of structural risk minimization. It seeks the best compromise between the complexity of the model and the learning ability according to the limited sample information to obtain the best generalization ability. The SVM model maps instances to points in space and then distinguishes instances of different categories at obvious intervals. SVM can not only deal with linear classification but can also introduce kernel functions to classify nonlinear problems. If a dichotomy is taken as an example, the classification diagram is shown in Figure 11.

Among them, the classification decision function is as follows:

$$f(x) = \omega^T x + b. \tag{3}$$

To improve the classification effect, the optimal classification plane can be obtained by maximizing the classification interval.

$$\begin{aligned} &\max \frac{1}{\|\omega\|}, \\ &\text{s.t.} \quad y_i\left(\omega^T x_i + b\right) \geq 1 \quad i = 1, \ldots, n. \end{aligned} \tag{4}$$

Equation (4) can be solved by introducing the duality principle of Lagrange's theorem.

$$f(\omega, b, a) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n} a_i\left[y_i\left(\omega^T x_i + b\right) - 1\right]. \tag{5}$$

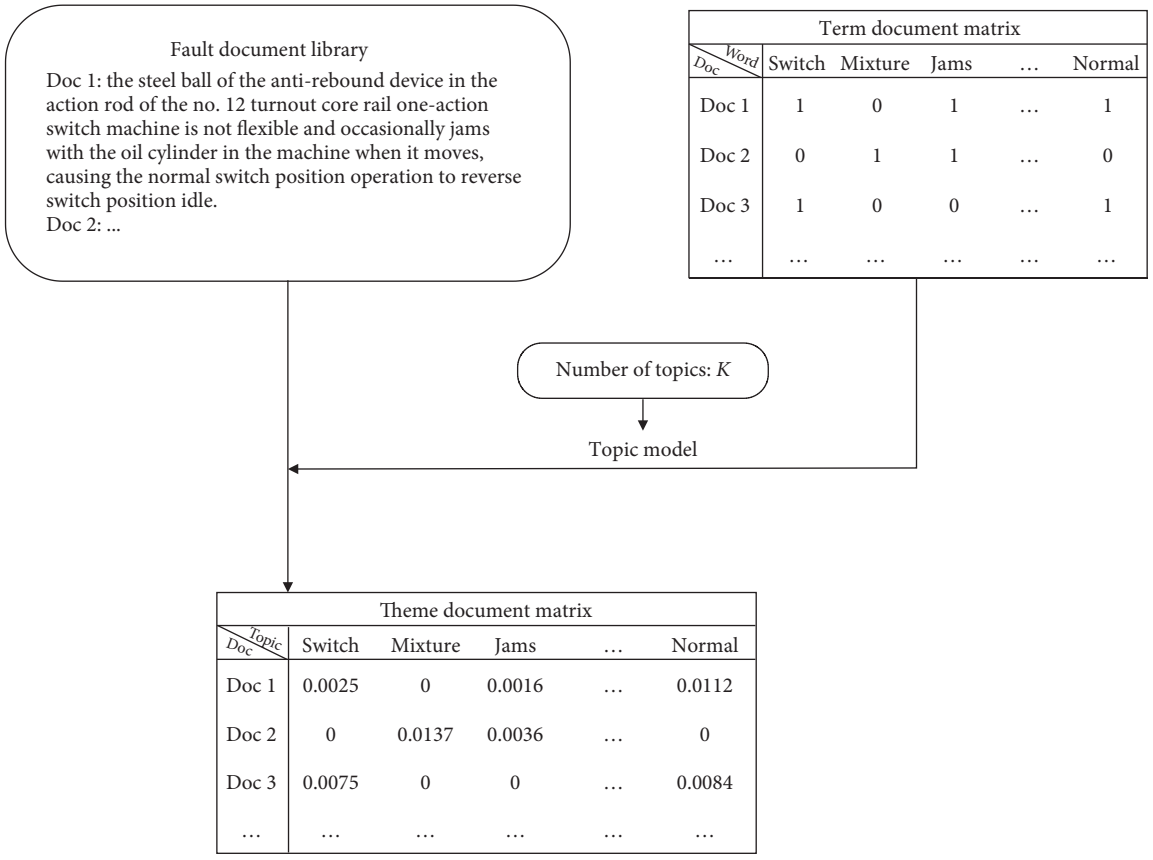

Figure 10: Expression process of fault document in topic space.

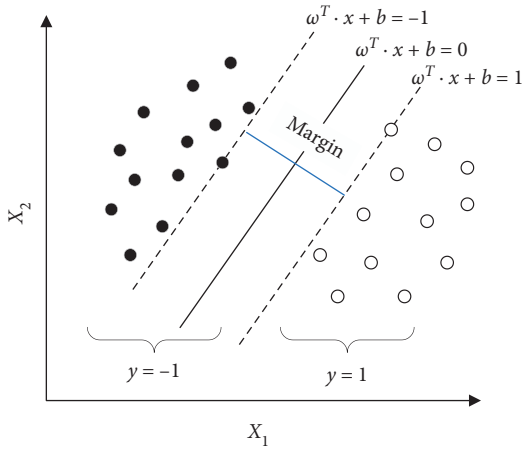

Figure 11: SVM linear binary classification diagram.

In practice, there will be some data points that deviate from the decision plane, so the relaxation variable ξ_i is introduced here. Therefore, equation (5) can be transformed into the following:

$$f(\omega, b, a, \xi, \gamma) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n} a_i\left[y_i\left(\omega^T x_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\gamma_i\xi_i, \tag{6}$$

where C is the penalty factor, which is used to constrain the relaxation variable.

The partial derivative of equation (6) can be found and then rearranged.

$$f(x) = \left(\sum_{i=1}^{n} a_i y_i x_i\right)^T x + b = \sum_{i=1}^{n} a_i y_i \langle x_i, x_j\rangle + b, \tag{7}$$

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i, a_j, y_i, y_j\langle x_i, x_j\rangle, \\ \text{s.t.} \quad & 0 \le a_i \le C \quad i = 1, \ldots, n, \\ & \sum_{i=1}^{n} a_i y_i = 0. \end{aligned} \tag{8}$$

To obtain the value of vector a, the sequence minimum optimization (SMO) algorithm [30] is introduced. Knowing the value of a, the value of ω can be obtained. According to equation (9), b can be obtained.

$$y_s\left(\sum_{i=1}^{S} a_i y_i K(x_i, x_s) + b\right) = 1. \tag{9}$$

By transforming the data into data points and then mapping the corresponding points to the high-latitude space, the linear inseparability situation can be solved easily

to realize the classification of the data. However, nonlinear mapping in a high-dimensional space will greatly increase the number of calculations, so introducing the concept of a kernel function into the support vector machine can effectively reduce the number of calculations. In this study, the Gaussian kernel function was used to implicitly map its input into high-dimensional space to make it more effective for nonlinear classification.

The above mainly describes the situation of the support vector machine for a dichotomy, but in reality, most data points will have more than just two types. At present, there are two main ways to realize multiclassification in an SVM algorithm: the one-vs-rest method (OVR SVM) and the one-vs-one method (OVO SVM).

4.2. Fault Diagnosis of Railway Signal Equipment Based on the SVM Algorithm. In this study, a one-vs-one multiclassification method (OVO SVM) was used to train the model. The subject document matrix and the fault category label corresponding to the signal equipment fault document were used as the training data of the model and input into the classifier to obtain the corresponding SVM diagnostic model. For a new signal equipment fault document, the document was first preprocessed, then the preprocessed document was transformed into the topic feature space. The previously trained SVM diagnostic model was then input to diagnose the new fault so that the type of fault that occurred could be obtained to realize the fault diagnosis of the railway signal equipment. The corresponding diagnosis process of the SVM is shown in Figure 12.

5. Experimental Analysis

5.1. Evaluation Index of the SVM Algorithm. In the design process of the classifier, an index is usually needed to evaluate the performance of the classifier, which is not only beneficial to the subsequent optimization of the classifier but can also intuitively show the quality of the classification effect.

Precision is usually used to evaluate the performance of a classifier. The precision rate represents the proportion of samples that are positive to the samples predicted by the model to be positive. However, when the distribution ratios of different categories of sample data differ greatly, the precision rate cannot fully reflect the classification effect of a classifier. That is because when the data distribution is not balanced, the sample data of a small proportion category is easy to be wrongly classified into the sample of a large proportion category. Because there is such a small amount of data in a small-scale sample, even if the classification effect of the small-scale sample is poor, it will not have a great impact on the overall precision of the classifier. To make up for this shortcoming, the $F1$-measure was introduced in this study as another evaluation indicator to evaluate the effect of the classifier more comprehensively.

The F-measure is calculated by Precision and Recall. The precision rate refers to the proportion of correctly predicted positive samples to the total number of predicted positive samples. The precision rate can directly show the rate of accurate samples within the classification samples.

$$P = \text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{10}$$

In the above equation, TP is the number of samples correctly categorized into this class, and FP is the number of samples wrongly assigned to this class.

The recall rate refers to the percentage of the predicted positive samples in the actual positive samples.

$$R = \text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{11}$$

In equation (11), FN represents the number of samples that belong to this category but have been misclassified into another category.

When considering the classification effect of each failure separately, there will often be contradictions between Precision and Recall. The F-measure is the weighted average of Precision and Recall, so it can effectively solve the above contradiction. The expression for the F-measure is as follows:

$$F - measure = \frac{\left(\mu^2 + 1\right) \times \text{Precision}}{\mu^2 \times (\text{Precision} \times \text{Recall})}. \tag{12}$$

The following equation shows how the $F1$-measure used in this study was obtained, for when $\mu = 1$.

$$F1 - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}. \tag{13}$$

5.2. Effect Analysis of the Fault Diagnosis Experiment. To verify the proposed feature extraction and fault diagnosis algorithm, the fault diagnosis program was written using the Python programming language, and the corresponding experimental test and verification were performed on a PC configured with an Intel Core i7-6700HQ 2.60 GHz CPU and 16 GB memory.

To ensure that it was consistent with the actual fault diagnosis of the railway signaling and communications department, the 1000 fault text data samples used in the experiment were all taken from fault data recorded by the Lan-Xin high-speed railway signaling and communications department from 2017–2019.

The Naive Bayes (NB) algorithm, logistic regression (LR) algorithm, random forest (RF) algorithm, K-nearest neighbor (KNN) algorithm, and support vector machine (SVM) algorithm were selected to be used in the experiment. The widely used word-space model and the LDA topic model were used to extract the features of the data set, and then the two processed data sets were trained and tested. Through the Precision, Recall, $F1$-measure, and other indicators, the

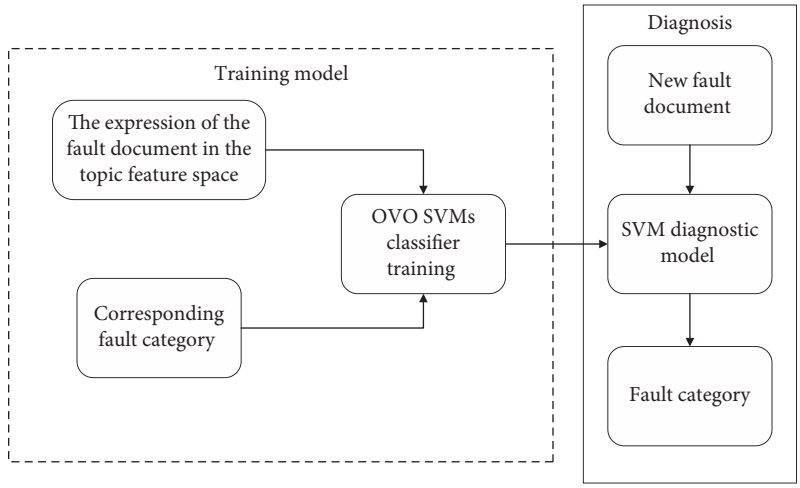

Figure 12: Fault diagnosis model training and new fault diagnosis process.

diagnosis effect was evaluated, analyzed, and compared to verify the impact of the LDA topic model on the performance of the fault diagnosis. To prevent overfitting problems, this experiment randomly used 70% of the data set to train the classifier, and the remaining 30% was used as the test set.

5.2.1. Experiment on Spatial Classification of Word Features. The fault text data was expressed in the traditional vector space model. Then the processed data was input into the classifier for training. The classification effect is shown in Table 3.

It can be seen from Table 3 that under the traditional word feature space method, the KNN algorithm had the best classification effect, while the RF algorithm in the ensemble classifier was the least effective.

5.2.2. Experiment of Topic Feature Space Classification. After using the LDA topic model to extract fault features from the same original fault data, the classifier was trained and tested again. The classification results are shown in Table 4.

As Table 4 shows, after the LDA topic model extracted the fault record text features and transformed them into the topic feature space, the classification indexes of most classifiers were improved to a certain extent.

5.3. Experimental Analysis of Fault Diagnosis Algorithm. Through a comparative analysis of the above experiments, it can be seen that the method of fault diagnosis of railway signal equipment based on the LDA and SVM models is better than the established method of combining word feature space with various classifiers. It shows that the method based on machine learning and natural language processing proposed in this paper has certain advantages in the fault diagnosis of railway signal equipment and can provide a certain reference value for railway signal equipment fault diagnosis.

Table 3: Classification effect of word feature space.

Method	Precision	Recall	*F*1-measure
LR	0.77	0.84	0.80
NB	0.76	0.85	0.80
RF	0.66	0.75	0.68
KNN	0.82	0.70	0.72
SVM	0.80	0.87	0.83

Table 4: Classification effect of topic feature space.

Method	Precision	Recall	*F*1-measure
LDA + LR	0.81	0.84	0.81
LDA + NB	0.77	0.74	0.75
LDA + RF	0.75	0.82	0.78
LDA + KNN	0.80	0.81	0.80
LDA + SVM	0.84	0.87	0.83

6. Conclusions

In this paper, based on the characteristics of the text data of fault records of railway signal equipment from the Lan-Xin high-speed railway, a corresponding fault feature vocabulary in the field of railway signal equipment was constructed. The LDA topic model was used to extract the fault features. Then the fault record text was transformed into the topic feature space and compared with the commonly used word feature space. Subsequently, an SVM diagnostic device was constructed based on this to diagnose the fault. Then it was compared with the NB, LR, RF, and KNN algorithms. The accuracy and effectiveness of the proposed model were verified by field data experiments. It is proved that the model can help the field personnel to quickly diagnose and maintain the railway signal equipment failure and has certain guiding significance.

In future research, we will try to process the text records for a single type of signal equipment failure and test the influence of different classification algorithms on the corresponding text evaluation indicators to achieve further

diagnoses of certain types of signal equipment failure. At the same time, this method is also suitable for a large number of textual information on failure phenomena recorded in a variety of other industrial production activities and has a wide range of application prospects.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of China (51967010) and the Key Program of Natural Science Foundation of Gansu Province, China (20JR5RA428) (20JR10RA218).

References

[1] Y. Zhao and T. H. Xun, "Fault diagnosis system for vehicle onboard equipment of high-speed railway," *Journal of the China Railway Society*, vol. 37, no. 8, pp. 53–59, Aug. 2015.

[2] L. B. Yang, P. Li, R. Xue, X. N. Ma, Y. H. Wu, and D. Zou, "Intelligent classification of faults of railway signal equipment based on imbalanced text data mining," *Journal of the China Railway Society*, vol. 40, no. 2, pp. 59–66, 2018.

[3] Z. W. Zhong, T. Tang, and F. Wang, "Research on fault feature extraction and diagnosis of railway switches based on PLSA and SVM," *Journal of the China Railway Society*, vol. 40, no. 7, pp. 80–87, 2018.

[4] W. ShangGuan, Y. H. Yuan, J. Wang, and F. W. Hu, "Research of fault feature extraction and diagnosis method for CTCS onboard equipment(OBE) based on labeled-LDA," *Journal of the China Railway Society*, vol. 41, no. 8, pp. 56–66, 2018.

[5] Y. Ren and J. Z. Pan, "Optimising ontology stream reasoning with truth maintenance system," in *Proceedings of the 20th AMC CIKM 2011*, pp. 831–836, Glasgow, UK, October 2011.

[6] M. Hamadache, S. Dutta, O. Olaby et al., "On the fault detection and diagnosis of railway Switch and crossing systems: an overview," *Applied Sciences*, vol. 9, no. 23, 2019.

[7] Z. J. Zhao, X. K. Li, Q. Guo, K. Yang, and J. G. Liu, "Evolution properties of complex networks in terms of the LDA," *Journal of University of Electronic Science and Technology of China*, vol. 48, no. 6, pp. 931–938, 2019.

[8] Y. Peng, C. X. Wan, T. J. Jiang et al., "Extracting product aspects and user opinions based on semantic constrained LDA model," *Journal of Software*, vol. 28, no. 3, pp. 676–693, 2016.

[9] Z. Liu, T. Qin, K.-J. Chen, and Y. Li, "Collaboratively modeling and embedding of latent topics for short texts," *IEEE Access*, vol. 8, pp. 99141–99153, 2020.

[10] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimedia Tools and Applications*, vol. 80, pp. 1–31, 2021.

[11] Z. Liu, C. Y. Liu, B. Xia, and T. Li, "Multiple relational topic modeling for noisy short texts," *International Journal of Software Engineering and Knowledge Engineering*, vol. 28, no. 11n12, p. 16, 2019.

[12] G. Xu, X. Wu, H. Yao, F. Li, and Z. Yu, "Research on topic recognition of network sensitive information based on SW-LDA model," *IEEE Access*, vol. 7, pp. 21527–21538, 2019.

[13] L. Shi, G. Song, G. Cheng, and X. Liu, "A user-based aggregation topic model for understanding user's preference and intention in social network," *Neurocomputing*, vol. 413, pp. 1–13, 2020.

[14] Z. Zhao, H. N. Koutsopoulos, and J. H. Zhao, "Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model," *Transportation Research Part C*, vol. 116, pp. 1423–1436, 2020.

[15] S. S. Erzurumlu and D. A. Pachamanova, "Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations," *Technological Forecasting and Social Change*, vol. 156, 2020.

[16] W. T. Zhou, H. B. Wang, H. G. Sun, and T. L. Sun, "A method of short text representation based on the feature probability embedded vector," *Sensors*, vol. 19, no. 17, p. 3728, 2019.

[17] X. Fu, X. Sun, H. Wu, L. Cui, and J. Z. Huang, "Weakly supervised topic sentiment joint model with word embeddings," *Knowledge-Based Systems*, vol. 147, pp. 43–54, 2018.

[18] G. H. Feng and W. Zhen, "Review of Chinese automatic word segmentation," *Library and Information Service*, vol. 55, no. 2, pp. 41–45, 2011.

[19] Q. Liu and H. B. Jia, "A view of Chinese word automatic segmentation research in the Chinese information disposal," *Computer Engineering and Applications*, vol. 3, pp. 175–177+182, 2006.

[20] T. Williams and J. Betak, "A comparison of LSA and LDA for the analysis of railroad accident text," *Procedia Computer Science*, vol. 130, pp. 98–102, 2018.

[21] P. Celard, A. S. Vieira, E. L. Iglesias, and L. Borrajo, "LDA filter: a latent dirichlet allocation preprocess method for weka," *PLoS One*, vol. 15, no. 11, p. e0241701, 2020.

[22] S. Xiong, K. Wang, D. Ji, and B. Wang, "A short text sentiment-topic model for product reviews," *Neurocomputing*, vol. 297, pp. 94–102, 2018.

[23] L. Hagen, "Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models?" *Information Processing & Management*, vol. 54, no. 6, pp. 1292–1307, 2018.

[24] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Systems With Applications*, vol. 127, pp. 256–271, 2019.

[25] R. Li, H. P. Zhang, Y. P. Zhao, and J. Y. Shang, "Automatic text summarization research based on topic model and information entropy," *Computer Science*, vol. 41, no. S2, pp. 298–300+332, 2014.

[26] K. S. Gyamfi, J. Brusey, A. Hunt, and E. Gaura, "A dynamic linear model for heteroscedastic LDA under class imbalance," *Neurocomputing*, vol. 343, pp. 65–75, 2019.

[27] M. Juszczyk, "On the search of models for early cost estimates of bridges: an SVM-based approach," *Buildings*, vol. 10, no. 1, p. 2, 2019.

[28] D. C. Toledo-Pérez, J. Rodríguez-Reséndiz, R. A. Gómez-Loenzo, and J. C. Jauregui-Correa, "Support vector machine-based EMG signal classification techniques: a review," *Applied Sciences*, vol. 9, no. 20, p. 4402, 2019.

[29] S. Yadavar Nikravesh, H. Rezaie, M. Kilpatrik, and H. Taheri, "Intelligent fault diagnosis of bearings based on energy levels in frequency bands using wavelet and support vector machines (SVM)," *Journal of Manufacturing and Materials Processing*, vol. 3, no. 1, p. 11, 2019.

[30] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2014.