Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7882396, 12 pages
https://doi.org/10.1155/2022/7882396

Research Article

Railway Fault Text Clustering Method Using an Improved Dirichlet Multinomial Mixture Model

Ni Yang and Youpeng Zhang

School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Correspondence should be addressed to Youpeng Zhang; zhangyp@mail.lzjtu.cn

Received 15 April 2022; Revised 23 May 2022; Accepted 27 May 2022; Published 4 July 2022

Academic Editor: Naeem Jan

Railway signal equipment fault data (RSEFD) are one of the issues with in-depth traffic big data analysis throughout the life cycle of intelligent transportation. In the course of daily operation and maintenance, the railway electrical maintenance department records equipment malfunction information in a natural language. The data have the characteristics of strong professionalism, short text, unbalanced category, and low efficiency of manual analysis and processing. How to effectively mine the information contained in these fault texts to provide help for on-site operation and maintenance plays an important role. Therefore, we propose a railway fault text clustering method using an improved Dirichlet multinomial mixture model called ICH-GSDMM. In this method, first, the railway signal terminology thesaurus is established to overcome the inaccurate problem of RSEFD segmentation. Second, the traditional Chi square statistics is improved to overcome the learning difficulties caused by the imbalance of RSEFD. Finally, the Gibbs sampling algorithm for Dirichlet multinomial mixture model (GSDMM) is modified using an improved chi-square statistical method (ICH) to overcome the symmetry problem of the word Dirichlet prior parameters in the traditional GSDMM. Compared to the traditional GSDMM model and the GSDMM model based on chi-square statistics (CH-GSDMM), the quantitative experimental results show that the GSDMM model based on improved chi-square statistics (ICH-GSDMM internal)'s evaluation index of clustering performance has greatly improved, and its external evaluation indices are also the best, with the exception of external index NMI of data set DS2. Simultaneously, the diagnostic accuracy of a select few categories in RSEFD has considerably improved, demonstrating its efficacy.

1. Introduction

The intelligent transportation system (ITS) is the development trend of transportation system in the future and it has received more and more attention. Depth analysis of traffic big data in the whole life cycle is becoming one of several scientific and technical problems in China's intelligent transportation, and at present, it is in a primary stage where the data is not wide enough and the application is not deep enough, and it has the problem that data integration and intelligence needs to be further improved. How to fully dig the value of massive data covering the entire life cycle of the transportation field has become the basis research and it has promoted the construction of a new generation of intelligent transportation systems [1]. Railway signal fault data (RSEFD) are one part of the massive data of the whole life cycle of the transportation field and it has received more and more attention.

Railway signal equipment generally refers to track circuits, signals, turnouts, and other equipment related to train operation, and these equipment are the basis for ensuring the safe operation of trains. In the daily operation process, maintenance personnel record the fault phenomenon, the handling process of equipment failure and fault diagnosis results in a natural language, and store the fault data in paper or electronic files in text. With the increase of railway mileage and operation, a large number of RSEFD have been accumulated. These data are stored in unstructured and textual form for a long time, and it is not conducive to computer processing and understanding [2]. Equipment

maintenance workers must frequently learn from the processing experience of a significant number of existing equipment fault data, as well as manual inquiry and analysis of this fault data, during normal maintenance of railway signal equipment. This results in low data processing efficiency and low intelligence of data information [3, 4].We effectively reduce the search space, improve the discovery efficiency, and mine a large amount of valuable fault identification and diagnosis information contained in the fault text [5, 6] as well as established the association between fault feature words and fault classes that will make fault identification effective and similar situations handling easy in the future [7]. Railway personnel manually classify the severity and domain reflected by the textual semantics of railway faults based on professional knowledge [8]. Due to the unstructured structure of railway text data and the irregularity and randomness of personnel records [3], it is currently a challenge to extract accurate fault information from unstructured natural language.

The topic model is a traditional text clustering method, which can well mine the semantic information of text and is widely used. As the most popular topic model [9], LDA is often used for text clustering, and it is successfully applied to long text clustering and the effect is successful. Short texts tend to have fewer words and data sparsity. Due to the lack of repeated words in short texts, it is a challenging task for the traditional LDA topic model to screen relevant feature words. Meanwhile, in short texts, the context is very limited, and semantic-based feature word extraction is challenging. When traditional topic modeling techniques are applied to RSEFD, it is necessary to consider the characteristics of short texts, and feature extraction algorithms that copy long texts are often ineffective.

GSDMM [10] can automatically deduce the number of clusters and it has a good balance between the completeness and homogeneity of clustering results, as well as a fast convergence speed, which is more effective than LDA to extract hidden topics from short texts [11]. The GSDMM model assumes that the parameter of words Dirichlet prior distribution is symmetric, that is, the same Dirichlet prior distribution is given to all words, and all words are treated equally when the model is generated. In practice, different words should have different clustering effects on topics, GSDMM should consider the influence of global weighted metrics for each word [12], and the parameters of Dirichlet prior distribution of each word should be different.

To address the challenges posed by the symmetric assumption of the parameter of words Dirichlet prior distribution of the GSDMM model, chi-square statistics is introduced. Chi-square statistic tests the significance of the relationship between the value of a variable and that class [13]. The importance of different words to different classes can be well distinguished by chi square statistic (CHI). The larger the chi-square statistic value of a feature item in a specific class, the more representative the word is for that class. Chi-square statistics have greatly improved the sparseness of feature words in short text datasets. However, chi-square statistics also have shortcomings. The traditional chi-square statistical algorithm does not take the uniform distribution of feature words within the class into account and ignores some features that rarely appear in the specified category but can well represent this category [14–16]. The imbalance of fault data categories affects the performance of feature extraction algorithms and also brings serious difficulties to most clustering models and classifier learning algorithms that assume a relatively balanced data distribution [7].

To solve the above problems, in order to further improve the mining quality of the hidden information of the fault text and improve the clustering effect of the railway signal fault text, this paper proposes an ICH-GSDMM model for railway fault text clustering, and the main contributions are summarized as follows:

(1) A professional word segmentation dictionary in the field of railway signal is constructed. The natural language of signal fault text is highly specialized and general text segmentation tools are not effective for some professional vocabulary segmentation. The establishment of this dictionary effectively improves the word segmentation accuracy of signal fault text and provides a good basic environment for feature words to better represent text semantics and improve text clustering effect.

(2) A feature word extraction method based on prior knowledge of improved chi-square statistics is proposed. This method filters out the feature words of each category based on the relationship between the feature words and the categories, which effectively alleviates the problem of loose topics in short texts and greatly improves the problem of inaccurate feature word extraction caused by imbalanced data categories.

(3) An ICH-GSDMM model based on prior knowledge of railway domain is constructed. By changing the weight of Dirichlet prior distribution of each word in the GSDMM model, the model improves the semantic balance of the fault text representation vector generated by the GSDMM model and improves the text clustering effect.

The remainder of this paper is organized as follows. Section 2 reviews the literature on topic models and chi-square statistics. Section 3 explains feature word extraction algorithm with improved chi-square statistics. Section 4 elaborates the GSDMM and the ICH-GSDMM model. Section 5 is the experimental data and analysis. Section 6 summarizes the paper and proposes future work.

2. Related Work

How to remove hidden fault information from fault text for clustering and equipment fault type identification is the main work carried out in the field of railway signal fault text earlier. For example, the authors of [4] used the TF-IDF algorithm for feature word extraction, and then integrated multiple classifiers based on voting to achieve fault text classification learning. The authors of [17] applied Word2vec

to generate word vectors and the SMOTE algorithm to balance the amount of data, and finally used convolutional neural networks to automatically classify faulty texts. The authors of [18] put forward a method for fault text classification based on Word2vec and parallel convolutional neural networks. Based on the high-speed rail signal equipment fault text, the authors of [3, 19] adopted the PLSA model and the labeled-LDA topic model for feature extraction and fault text clustering respectively, so as to realize fault diagnosis of on-board equipment in high-speed rail signaling systems. In the study of [7], to classify the problematic text, the authors presented the syntactic feature extraction approach of enhanced chi-square statistics and the semantic feature extraction method of LDA topic model based on prior knowledge. The above method usually represents the text as a vector by calculating the word frequency or semantic information of the feature words in the fault text and then calculates the similarity and realizes clustering or classification.

Topic modeling approaches make it possible to cluster enormous amounts of unlabeled data efficiently. It is an unsupervised machine learning model that belongs to the soft clustering method and can effectively extract semantic information in the text to mine the topic of clustered text. Each text is supposed to be a mixture of topics in the LDA model [20], with each topic consisting of a set of connected words that usually communicate some semantic information [9, 21]. Since the railway signal fault text belongs to the short text domain, there are few repeated words in the short text, and the data are sparse, which lead to the unsatisfactory estimation of the topic distribution of the text and the topic distribution of words by LDA. The GSDMM proposed by the authors of [10] is more suitable for short text clustering. Compared with other topic clustering methods, the short text topic vectors generated by GSDMM are of better quality, the clustering results have good integrity and homogeneity, and the convergence speed is fast, and it can also deal with the sparse and high-dimensional problems of short texts. The GSDMM model is the Dirichlet multinomial mixture (DMM) model based on the folded Gibbs sampling algorithm, which assumes that each document can only be represented by one topic. The authors of [22] adopted the GSDMM method for short text clustering in the field of web services, and the performance study showed that GSDMM is a more effective clustering method compared to other traditional topic modeling methods. The authors of [23] first used the GSDMM topic model to generate the corresponding topic vector of the text, and then applied the AGNES algorithm to analyze the clustering effect of the topic vector. The research results showed that the GSDMM topic model method has better clustering quality for the service text. The authors of [24] proposed a FGSDMM + algorithm, which uses multiple runs of the folded Gibbs sampling algorithm to complete online text clustering. Compared with the GSDMM and FGSDMM algorithms, the final clustering performance shows that the FGSDMM + algorithm has better data clustering performance. The authors of [25] put forward an adaptive Dirichlet multinomial mixture clustering model (e-GSDMM), which utilizes a hyperparameter tuning algorithm to automatically capture temporal dynamics to obtain the temporal variation of topics and word distributions for short texts, the clustering results show that e-GSDMM outperforms existing GSDMM methods on short text streaming data. In summary, at present, there are few improvement studies on the assumption that the word Dirichlet prior distribution is symmetrical in the GSDMM model.

The larger the chi-square statistic value of a feature item in a specific class, the more representative the word is for that class. Chi-square statistics are often used for feature selection [26, 27]. Because basic chi-square statistics are insufficient, several researchers have improved them. The authors of [15] proposed a modified chi-square statistics for feature selection approach and confirmed its efficacy based on the word frequency of feature items and their distribution features between and among classes. Aiming at the problem of missing attributes in some classes in chi-square statistics, the authors of [28] balanced the screening of the number of feature words in each class by improving the chi-square statistical algorithm and combines the SVM classifier to modify the performance of the Arabic text classification model. The above research on chi-square statistics in text classification models also illustrates the effectiveness of chi-square statistics in the field of text classification. For above considerations, in this paper, a research on railway signal fault text clustering based on ICH-GSDMM is carried out.

3. Feature Extraction Based on Improved Chi-Square Statistics

The purpose of chi-square statistics reference is to effectively extract the fault feature words of each category and reduce the impact of fault category imbalance on text clustering.

3.1. Chi-Square Statistics. Chi-square statistics (CH) is used to measure the degree of correlation between words and classes, and it is assumed that words w_i and c_i classes conform to a χ^2 distribution with a first degree of freedom. The higher the χ^2 statistic value of the entry w_i for a certain category c_i, the greater the correlation between it and the category, and the smaller the independence. The chi-square statistic is defined as [7]

$$\chi^2(w_i, c_i) = \frac{N\left[f(w_i, c_i) f(\overline{w}_i, \overline{c}_i) - f(w_i, \overline{c}_i) f(\overline{w}_i, c_i)\right]^2}{f(w_i) f(\overline{w}_i) f(c_i) f(\overline{c}_i)}, \tag{1}$$

where N is the number of documents in the corpus, $\overline{w}_i$ indicates that the word w_i is not included, $\overline{c}_i$ indicates other categories except class c_i in the corpus, $f(\cdot, \cdot)$ shows the relevance between the word w_i and class c_i, $f(w_i)$ indicates the number of texts in the corpus that contain the word w_i, $f(\overline{w}_i)$ indicates the number of texts in the corpus that does not contain the word w_i, $f(c_i)$ indicates the number of texts in the corpus that belong to class c_i, and $f(\overline{c}_i)$ indicates the number of texts in the corpus that do not belong to class c_i.

3.2. Improved Chi-Square Statistics. We refer to the class with a small number of texts as the minority class, and the class with more texts as the majority class, for clarity. For traditional chi-square statistics, only the frequency of documents containing feature words is considered, and the frequency of each feature word in these documents is not considered, which has disadvantages for corpora with uneven data distribution. The notion of frequency is presented to overcome the problem of unreliable feature word extraction due to the tiny amount of text contained in the minority class. The ideas of interclass concentration and intraclass dispersion are developed to overcome the problem that standard chi-square statistics increases the weight of feature words that appear less frequently in this class but commonly exist in other classes [16].

To facilitate understanding, we define K as the number of categories of a corpus, and a category C_i $(1 \le i \le K)$ contains text $d_{i1}, \ldots, d_{ij}, \ldots, d_{iM}$ $(1 \le j \le M)$ documents. The document frequency d_i^t of the feature word t appearing in the category C_i is defined as the intraclass dispersion, df_{ij}^t is the frequency of the feature word t appearing in the text d_{ij}, and cf_i^t is the frequency of the feature word t appearing in the category C_i, which is calculated as follows formula:

$$cf_i^t = \sqrt{\sum_{j=1}^{M} \left(df_{ij}^t\right)^2}, \tag{2}$$

where cf_t is the mean value of cf_i^t under all categories and the calculation is as follows:

$$cf_t = \frac{\sum_{i=1}^{K} cf_i^t}{K}, \tag{3}$$

where tf_i^t is the interclass concentration of the feature word t in the category C_i, and the calculation is as follows:

$$tf_i^t = \frac{\left(d_i^t - cf_t\right)^2}{cf_t}. \tag{4}$$

The calculation of improved chi-square statistics (ICH) is as follows:

$$\chi_{\text{new}}^2\left(w_i, c_j\right) = \chi^2\left(w_i, c_j\right) \times cf_i^t \times \mathrm{d}_i^t \times tf_i^t. \tag{5}$$

3.3. Feature Word Extraction. This paper first selects a fixed number of words as important feature words representing category according to the ICH. This filtering method effectively improves the feature words extraction quality of minority class and reduces the clustering problem due to class imbalance in the corpus. We define the improved chi-square statistic value of feature words as the ICH value, and the traditional chi-square statistic value of the feature words as the CHI value.

The feature word extraction method based on ICH feature selection is as in Algorithm 1. The RSEFD set S, the fault term dictionary Ω, the fault category set C, and the threshold γ is the number of important words in each category.

Algorithm 1 first initializes five empty sets, FS is the corpus set, which is used for the word set after data preprocessing, FI is the ICH value set, FI′ is the normalized FI set, Fw_c is the priori ICH value set, and FS′ is the important feature word set. (Line1-2). According to the fault term dictionary Ω, the corpus set FS is obtained after preprocessing the RSEFD set S, such as word segmentation and remove stop words (line3-4). Then calculate ICH values for all words and each category in the corpus set FS according to formula (5), and store them in the ICH value set FI (line6–9). In order to facilitate the comparison of the relationship between different fault feature words and different categories, the ICH value of each word in the set FI is normalized according to the following formula (line 10):

$$\chi^2\left(w_i, c_j\right) = \frac{\chi^2\left(w_i, c_j\right)}{\sum_{i=1}^{K} \chi^2\left(w_i, c_j\right)}, \tag{6}$$

where K is the number of categories in the RSEFD set S, w_i is a feature word in the corpus set FS, and c_j $(1 \le j \le K)$ is a category in the maintenance data set S. Next, FI′ is filtered according to the threshold γ to obtain the priori ICH value set Fw_c (line11–13). Finally, the important feature word set FS′ is obtained according to the priori ICH value set Fw_c (line14-15).

4. Clustering Algorithms

This section first introduces the traditional GSDMM model and its implementation algorithm and then explains the ICH-GSDMM model proposed in the text.

4.1. GSDMM Model. GSDMM is a DMM model with the folded Gibbs sampling algorithm, and it is a probabilistic generative unsupervised model. Under the assumption of one-to-one correspondence between topics and documents, GSDMM adopts an iterative Gibbs sampling algorithm to approximate the model, and finally generates the topic distribution of documents. Figure 1 shows a graphical representation of the simulated process of DMM generating documents.

In the DMM model, α is the topic Dirichlet prior distribution parameter, β is the word Dirichlet prior distribution parameter, θ is the topic distribution matrix of the document, φ is the topic distribution matrix of the word, θ and φ satisfies

$$\begin{aligned} \theta|\alpha &\sim Dir\,(\alpha), \\ \varphi K|\beta &\sim Dir\,(\beta), \end{aligned} \tag{7}$$

where $\theta_{k,d}$ is the probability distribution of document d on topic k, and all topic distributions of the same document d satisfies

$$\sum_{k=1}^{K} \theta_{k,d} = 1, \tag{8}$$

where $\varphi_{k,w}$ is the probability distribution of word w on topic k, and the topic distribution of all words w in the same document satisfies

$$\sum_{w=1}^{V} \varphi_{k,w} = 1. \tag{9}$$

Input: Maintenance data set S, fault term dictionary Ω
fault class set C, Threshold γ
Output: feature word set FS′, Priori chi square set Fw_c
begin
(1) Initialize the parameters
(2) FS = Φ, FS′ = Φ, FI = Φ, FI′ = Φ, Fw_c = Φ
(3) **for** $si \in$ S **do**
(4) FS = FS ∪ Word Set by word segmentation in si according to Ω
end
(6) **for** $w_i \in$ FS **do**
(7) **for** $c_j \in C$ **do**
(8) FI_i = compute the $\chi^2(w_i, c_j)$ by formula (5)
end
(9) FI = FI ∪ FI_i
end
(10) FI′ = Normalization of FI by formula (6)
(11) **for** $c_j \in C$ **do**
(12) **for** $w_i \in$ FI′ **do**
(13) Fw_c = Rank $(\chi'^2(w_i, c_j), c_j, \gamma)$
end
end
(14) **for** $w_i \in$ Fw_c **do**
(15) FS′ = FS′ ∪ w_i
end
end

Algorithm 1: ICH feature selection.

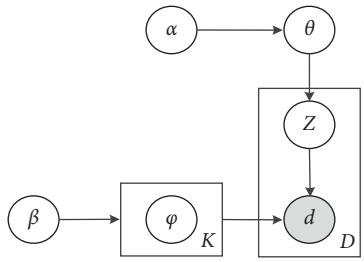

Figure 1: DMM graphical model.

The topic distribution of each document d obeys the following:

$$Z_d \sim \text{Mult}(\theta) \cdots d = 1, \ldots, D. \tag{10}$$

The process of document generation by DMM model can be described as follows: it first selects a mixed cluster k from formula (8). Then, it uses different algorithms to solve the model and finally get the probability that a topic k generates a document d as follows:

$$p(\mathrm{d}|z = k) = \prod_{w \in d} p(w|z = k). \tag{11}$$

GSDMM is an approximate solution algorithm model of the folded Gibbs sampling of DMM model. The approximate model of Gibbs sampling algorithm obtains θ and φ by continuously sampling different topics of a word according to formula (12), and finally we deduce the topic of each document.

$$p\left(z_{di} = k|\vec{z}_{-di}, \vec{d}\right) = \frac{m_{z,-di} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w_i \in d_i} \prod_{j=1}^{N_{d_i}^{w_i}} \left(n_{z,-d_i}^{w_i} + \beta + j - 1\right)}{\prod_{i=1}^{N_{d_i}} \left(n_{z,-d_i} + V\beta + i - 1\right)}. \tag{12}$$

4.2. ICH-GSDMM Model. In this section, we explain the ICH-GSDMM model suggested in this paper, which introduces frequency, intraclass concentration, and interclass dispersion in the traditional chi-square statistics. First, the important feature words W_{imp} of each classification are screened out according to the threshold γ, and then, the ICH value of the important feature words of each category is mapped to $[\lambda_1, \lambda_2]$, and used as the Dirichlet prior distribution of these important words, namely, β_1', and the Dirichlet prior distribution β_2' of the remaining feature words are all as λ_1.

In the ICH-GSDMM model, the probability of document d selecting cluster k is as follows:

$$p'\left(z_{di} = k|\vec{z}_{-di}, \vec{d}\right) = \frac{m_{z,-di} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w_i \in d_i} \prod_{j=1}^{N_{d_i}^{w_i}} \left(n_{z,-d_i}^{w_i} + \beta' j - 1\right)}{\prod_{i=1}^{N_{d_i}} \left(n_{z,-d_i} + V\beta' + i - 1\right)},$$

$$\beta' = \begin{cases} \beta_1' & w_i \in W_{imp} \\ \beta_2' & w_i \notin W_{imp} \end{cases},$$

$$\beta_1' = \lambda_1 + \frac{\lambda_2 - \lambda_1}{\chi_{\max}' - \chi_{\min}'}, \tag{13}$$

Table 1: Symbols in the ICH-GSDMM model.

D	Number of documents in the corpus
V	Number of words in the corpus
d_i	A document in the corpus
K	Assumed maximum number of clusters
N_{iter}	Number of iterations
α	Parameter of topic Dirichlet prior
β	Parameter of word Dirichlet prior $\vec{z}$ cluster labels of each document
$z_{\text{d}i}$	The cluster of document di
z_{cru}	A cluster label
z_{new}	A cluster label
$N_{\text{d}i}$	Number of words in document di
$N_{\text{z},-\text{d}i}$	Number of occurrences of word wi in cluster z without considering document di
$m_{\text{zd}i}$	Number of documents in cluster $z_{\text{d}i}$
$m_{\text{z},-\text{zd}i}$	Number of documents in cluster $z_{\text{d}i}$ Without considering document di
$n_{\text{zd}i}$	Number of words in cluster $z_{\text{d}i}$
$n_{\text{zd}i}^{wi}$	Frequency of word wi in cluster $z_{\text{d}i}$
$n_{\text{z},-\text{d}i}^{wi}$	Frequency of word wi in cluster z Without considering document di
$n_{\text{d}i}^{wi}$	Frequency of word wi in document di

where χ'max is the maximum value in the Fw_c set and χ'min is the minimum value in the Fw_c set.

Table 1 displays the symbols in the ICH-GSDMM model, and Algorithm 2 describes the main steps of the ICH-GSDMM model.

First, $n_{z_{\text{d}i}}$, $m_{z_{\text{d}i}}$, and $n_{z_{\text{d}i}}^{wi}$ are initialized by line 1. Then, the important feature word set FS′ and the priori chi square set Fw_c is obtained by calling algorithm 1. Next, the correction parameter β' of the Dirichlet prior distribution in the GSDMM model is obtained according to formula (13) by line3. The topic of each document in the corpus is then initialized (Lines 4–8). (line9–18) is the iterative calculation process of GSDMM based on the folded Gibbs sampling algorithm according to formula (13). Finally, the document-topic distribution matrix of the corpus is obtained according to the ICH-GSDMM model.

5. Experiments and Analysis

5.1. Evaluation Metrics. The evaluation indicators used to evaluate the performance of clustering algorithms can generally be divided into two categories: internal and external evaluations. Internal evaluation does not require ground-truth labels, and it evaluates the clustering effect by using some similarity measurement techniques to measure intraclass and interclass relationships. External evaluation requires ground truth labels, whether the clustering is reasonable is evaluated by analyzing the relationship between the clustering labels and the ground truth labels.

In our study, the internal evaluation indices adopt the silhouette coefficient (SC) [29] and the Davies–Bouldin coefficient (DBI) [30]. The external evaluation indices adopt normalized mutual information (NMI) [29], adjusted mutual information (AMI) [29], homogeneity (H) index [11], and integrity (C) [11].

5.1.1. Internal Evaluation

(1) Silhouette Coefficient. The silhouette coefficient (SC) is used to measure the separation distance between clusters. The formula for a single cluster SC is as follows:

$$SC_k = \frac{1}{N}\sum_{i=1}^{N}\frac{a_i - b_i}{\max(a_i, b_i)}, \tag{14}$$

where a_i is the average distance of element i from other elements in the same category and b_i is the average distance of elements that are closest to element i and belong to other different categories, N is the total number of elements in a cluster k.

The mean value of SC_k for each cluster k is the final silhouette coefficient score for all clusters with the following formula:

$$SC = \frac{1}{K}\sum_{k=1}^{K} SC_k. \tag{15}$$

The value of SC represents the quality of clustering performance, the higher the value, the better the clustering performance.

(2) Davidson Boding Coefficient (DBI). DBI calculates the distance between clusters and within clusters, and it is defined is as follows:

$$DBI = \frac{1}{N}\sum_{n=1}^{N}\max_{j\neq i}\left[\frac{\sigma_i + \sigma_j}{d(x_i, x_j)}\right], \tag{16}$$

where N is the number of categories of clusters, x_i and x_j are the ith and jth cluster centers, respectively, and σ_i and σ_j are the average distances from all points in the ith and jth clusters to the center point, respectively. DBI values reflect how similar texts are within the same and different clusters. The lower the DBI value, the better the clustering algorithm.

5.1.2. External Evaluation. The external evaluation indices NMI, AMI, and ARI all require ground truth labels and cluster labels.

(1) Normalized Mutual Information. Normalized mutual information (NMI) is defined as follows:

$$NMI(X, Y) = 2\frac{MI(X, Y)}{H(X) + H(Y)}, \tag{17}$$

where $X = \{x_1, x_2, \ldots, x_N\}$ is the cluster division after clustering and $Y = \{y_1, y_2, \ldots, y_N\}$ is the real category division. $H(X)$ and $H(Y)$ denote the entropy of X and Y, respectively, MI(X, Y) represents the mutual information calculation formula between X and Y.

(2) Adjust Mutual Information. Adjust mutual information (AMI) calculation formula is as follows:

$$AMI(X, Y) = \frac{MI(X, Y) - E\{MI(X, Y)\}}{(H(X) + H(Y)/2) - E\{MI(X, Y)\}}, \tag{18}$$

where $E(.)$ is the expectation of MI(X, Y).

(3) Homogeneity (H)

$$H = 1 - \frac{H(C|K)}{H(C)},$$

$$H(C|K) = -\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{n_{c,k}}{n}\log\left(\frac{n_{c,k}}{n_k}\right), \tag{19}$$

$$H(C) = -\sum_{c=1}^{C}\frac{n_c}{n}\log\left(\frac{n_c}{n_k}\right),$$

where $H(C)$ is the category division entropy, $H(C|K)$ is the conditional entropy of category division under the given clustering condition, n is the total number of texts in the corpus, n_c is the number of texts in category c, and n_k is the number of texts under cluster k. n_c, k represents the number of texts in class c which is divided into cluster k.

Homogeneity expresses the goal that each cluster contains elements of only one true group. A cluster is perfectly homogeneous if all elements in a cluster have the same external label.

(4) Completeness (C)

$$C = 1 - \frac{H(K|C)}{H(K)}. \tag{20}$$

The variable definitions of completeness are similar to homogeneity, and the definition of completeness is the conditional entropy of the cluster distribution given the external class labels. Completeness expresses the goal that all members with the same ground truth labels are assigned to one cluster.

5.1.3. Classification Correct Rate. To compare classification accuracy, we introduce the classification correct rate (CCR) [31]. The formula for CCR is as follows:

$$CCR = \frac{1}{n}\sum_{d=1}^{n}\delta\left(y_d, y_d'\right), \tag{21}$$

where n represents the total number of texts in the cluster and y'_d and y_d represent the predicted class label of document d and the highest-ranked label among the predicted class labels, respectively. $\delta(.)$ is an indicator variable, when classifying a multilabel data set, we define $\delta(y_d, y_d') = 1$ if y'_d is in y_d, and 0 otherwise. The larger the CCR value, the better the clustering performance. The introduction of classification accuracy can provide a good assessment of the performance of clustering models.

5.2. Experimental Data Set. The experimental data set DS1 selected in this paper is a Chinese data set, which is a RSEFD set collected by a railway company in China from 2016 to 2020, with a total of 1527 samples. In order to better test the clustering performance of the ICH-GSDMM model put forward in this paper, the English data set DS2 is also introduced. The data set DS2 is provided by https://github.com/pokarats/gsdmm, with a total of 20000 records. Table 2 shows examples of data set DS1 and DS2. Table 3 describes each fault category of data set DS1 and its proportion in the whole data set. It can be seen from Table 1 that the RSEFD set DS1 is a typical imbalanced data set. Track circuit fault (i.e., C2) and Switch fault (i.e., C5) are the majority, LKJ fault (i.e., C3) and Cab signal fault (i.e., C6) are the minority class. The classification accuracy of any fault category plays a key role in ensuring the safety and efficiency of the railway system. The data set DS2 contains 20 categories and each category contains 1000 samples.

5.3. Experimental Setup and Results. The experimental machine is configured with i7-10510u, 16.0GBRAM and win10.

Operating system and the program is written in Jupyter Notebook.

This section is described in two sections. The parameter settings for each topic modeling are described in the first section. In the second section, the clustering performances of GSDMM, CH-GSDMM, and ICH-GSDMM are evaluated and analyzed, respectively.

5.3.1. Parameter Setting. The β value of different prior Dirichlet distributions affects the performance of GSDMM. According to the literature [10], when the β value is [0.08, 0.1], the GSDMM model has high homogeneity and integrity, so this paper selects $\beta = 0.08$.

(1) In the GSDMM model, $\alpha = 0.1$ and $\beta = 0.08$.

(2) In the CH-GSDMM and ICH-GSDMM models, $\alpha = 0.1$ and $\lambda_1 = 0.08$, $\lambda_2 = 0.2$, and $\beta_2' = \lambda_1$.

(3) $K = 20$, $\gamma = 50$ in data set DS1. $K = 40$, $\gamma = 200$ in data set DS2. The number of iterations is 20, 40, and 60, and all experimental data are the mean values under different iterations.

5.3.2. Analysis and Discussion

(1) Internal Evaluation. Tables 4 and 5 indicate the SC, CH, and DBI results for the three topic models for 20, 40, and 60 number of clusters respectively. Table 6 depicts the mean values of internal evaluation results in DS1 and DS2. The mean values of SC, DBI, and CH of the ICH-GSDMM model are all better than those of the CH-GSDMM and GSDMM models, and the internal evaluation scores are improved more. For example, in the data set DS1, the SC score of ICH-GSDMM is 0.943, while the SC score of GSDMM model is 0.121. Compared with the traditional GSDMM model, the CH-GSDMM model or the ICH-GSDMM model can all significantly improve the internal evaluation performance of the clustering model.

(2) External Evaluation. The NMI, H, C, and DBI results for the three topic models for 20, 40, and 60 clusters are shown in Tables 7 and 8. Table 9 displays the mean values of

Input: K, α, β, Niter, D
Output: topic assignments to each document $\vec{Z}$
begin
(1) Initialize the parameters $m_{z_{di}}$, $m_{z_{di}}$ and n_{zdi}^{wi} as zeros for each cluster
(2) FS' ,Fw_c ← feature selection by Algorithm 1
(3) β' ← revised β by formula (13)
(4) for each document di ∈ D do
(5) z_{di} ← sample a cluster for di
(6) $m_{z_{di}} \leftarrow m_{z_{di}} + 1$ and $n_{z_{di}} \leftarrow n_{z_{di}} + N_{di}$
(7) for each word wi ∈ di do
(8) $n_{zdi}^{wi} \leftarrow n_{zdi}^{wi} + N_{di}^{wi}$
end
end
(9) for each iteration n in ∈ [1, N_{iter}] do
(10) for each document di ∈ D do
(11) z_{cru} ← Record the current cluster of di
(12) $m_{z_{cru}} \leftarrow m_{z_{cru}} - 1$ and $n_{z_{cru}} \leftarrow n_{z_{cru}} - N_{di}$
(13) for each word wi ∈ di do
(14) $n_{z_{cru}}^{wi} = n_{z_{cru}}^{wi} - N_{di}^{wi}$
end
(15) z_{new} ← sample a new cluster for di from formula (13)
(16) $m_{z_{new}} \leftarrow m_{z_{new}} + 1$ and $n_{z_{new}} \leftarrow n_{z_{new}} + N_{di}$
(17) for each word wi ∈ di do
(18) $n_{z_{new}}^{wi} = n_{z_{new}}^{wi} + N_{di}^{wi}$
end
end
end
(19) Return the result of topic distribution of each document $\vec{Z}$
end

Algorithm 2: The ICH-GSDMM algorithm.

Table 2: Examples of data sets DS1 and DS2.

Data set	Data description	Number
DS1	At 2 : 50 on November 5, 2019, 15608 train ran to k85 + 621 of the up line of the Ningxi line between CaiJiahe station and Bayuan station. Due to abnormal decoding of locomotive signal host, it stopped and affected one freight train	1527
DS2	How can I change Drupal's default menu strings without hacking the core files or using the string override plugin?	20000

Table 3: Fault classes and its percentage in data set DS1.

Index	Fault classes	Ratio (%)
C0	Interlocking fault	4.72
C1	ATP fault	16.76
C2	Track circuit fault	30.12
C3	LKJ fault	2.88
C4	Signal fault	10.87
C5	Switch fault	33.40
C6	Cab signal fault	1.24

external evaluation results in DS1 and DS2. The NMI and AMI score of the ICH-GSDMM model in the dataset DS1 are the same as the NMI and AMI score of the GSDMM model, and the scores of the rest external evaluation index H and C in the ICH-GSDMM model are the best among the three models. In the data set DS2, overall external evaluation result of ICH-GSDMM is better than CH-GSDMM and GSDMM models.

(3) CCR Analysis. CCR value is the average of the CCR values of each category in data set DS1 and DS2. Table 10 shows the results of the CCR scores in the data sets DS1 and DS2. It can be seen that the CCR score of the ICH-GSDMM model is the highest at 0.614, followed by CH-GSDMM and GSDMM.

The results of the CCR scores indicators for each class in the datasets DS1 and DS2 are shown in Figure 2.

In Figure 2(a), the data set DS1 contains 7 ground-truth labels, C0~C6. Because C1 (ATP fault) has little correlation with other classes, its CCR value reaches 1.0, which is better than CH-GSDMM and GSDMM models. Except for the C2 class, the CCR scores of other classes of the data set DS1 in the ICH-GSDMM model are better than those of the GSDMM and CH-GSDMM models. The reason for the lower CCR score of class C2 may be that C2 (Track circuit fault) is a basic ground equipment system for railway signals, which belongs to the majority classes in the data set DS1, and it has a greater correlation with class C2, C3, C4, and C5. Compared with the GSDMM model, the CCR scores of the

Table 4: Internal evaluation results after different iterations in DS1.

	GSDMM			CH-GSDMM			ICH-GSDMM		
	SC	DBI	CH	SC	DBI	CH	SC	DBI	CH
Niter = 20	0.114	1.413	241.387	0.941	0.478	6266.068	0.924	0.182	6564.016
Niter = 40	0.028	1.501	225.101	0.941	0.455	7304.858	0.946	0.364	6990.063
Niter = 60	0.220	1.597	300.127	0.940	0.212	8011.668	0.958	0.466	8516.038

Table 5: Internal evaluation results after different iterations in DS2.

	GSDMM			CH-GSDMM			ICH-GSDMM		
	SC	DBI	CH	SC	DBI	CH	SC	DBI	CH
Niter = 20	−0.041	2.120	0.912	0.579	0.783	3824.011	0.524	1.214	4148.330
Niter = 40	−0.048	2.104	1.300	0.431	1.106	3249.132	0.515	0.613	4178.312
Niter = 60	0.000	2.101	0.709	0.535	0.920	4430.288	0.528	0.981	4223.461

Table 6: Mean values of internal evaluation results.

	Topic model	SC	DBI	CH
DS1	GSDMM	0.121	1.504	255.538
	CH-GSDMM	0.941	0.382	7194.198
	ICH-GSDMM	0.943	0.338	7356.706
DS2	GSDMM	−0.030	2.108	0.974
	CH-GSDMM	0.515	0.936	3834.477
	ICH-GSDMM	0.522	0.936	4183.368

Table 7: External evaluation results after different iterations in DS1.

	GSDMM				CH-GSDMM				ICH-GSDMM			
	NMI	H	C	AMI	NMI	H	C	AMI	NMI	H	C	AMI
Niter = 20	0.753	0.638	0.705	0.750	0.580	0.630	0.538	0.576	0.712	0.793	0.647	0.709
Niter = 40	0.579	0.631	0.536	0.575	0.576	0.648	0.539	0.572	0.655	0.689	0.623	0.651
Niter = 60	0.642	0.654	0.612	0.638	0.595	0.639	0.591	0.591	0.610	0.606	0.615	0.606

Table 8: External evaluation results after different iterations in DS2.

	GSDMM				CH-GSDMM				ICH-GSDMM			
	NMI	H	C	AMI	NMI	H	C	AMI	NMI	H	C	AMI
Niter = 20	0.468	0.451	0.447	0.449	0.472	0.469	0.476	0.470	0.471	0.466	0.482	0.460
Niter = 40	0.481	0.467	0.466	0.466	0.440	0.462	0.449	0.438	0.464	0.459	0.478	0.462
Niter = 60	0.481	0.470	0.466	0.479	0.476	0.462	0.491	0.474	0.486	0.476	0.497	0.485

Table 9: Mean values of external evaluation results.

	Topic model	NMI	H	C	AMI
DS1	GSDMM	0.65	0.64	0.61	0.65
	CH-GSDMM	0.58	0.63	0.55	0.58
	ICH-GSDMM	0.65	0.69	0.62	0.65
DS2	GSDMM	0.47	0.46	0.46	0.4
	CH-GSDMM	0.46	0.46	0.47	0.46
	ICH-GSDMM	0.47	0.46	0.48	0.47

Table 10: CCR scores.

	Topic model	CCR
DS1	GSDMM	0.544
	CH-GSDMM	0.584
	ICH-GSDMM	0.614
DS2	GSDMM	0.753
	CH-GSDMM	0.763
	ICH-GSDMM	0.812

minority classes C0, C3, and C6 in the ICH-GSDMM model have been greatly improved. It can be seen that the prediction effect of the ICH-GSDMM model in the minority classes has been improved.

From the analysis of CCR performance of each class in Figure 2, it can be seen that the overall performance of the ICH-GSDMM model among the three models is still the best.

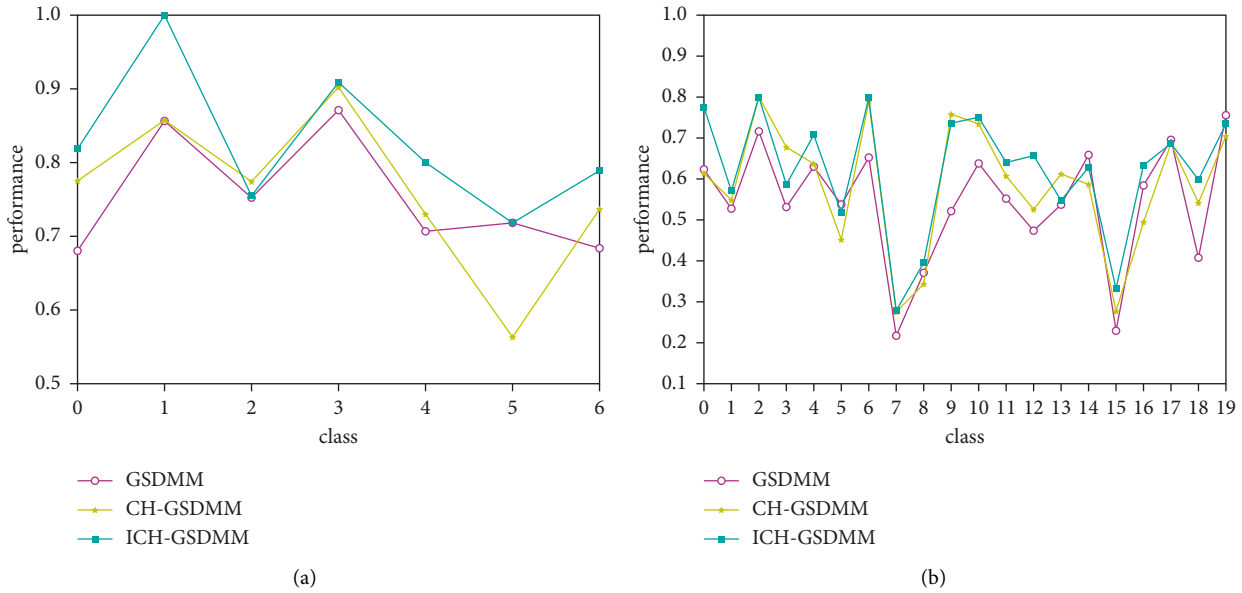

Figure 2: CCR scores for each category. (a) Data set DS1. (b) Data set DS2.

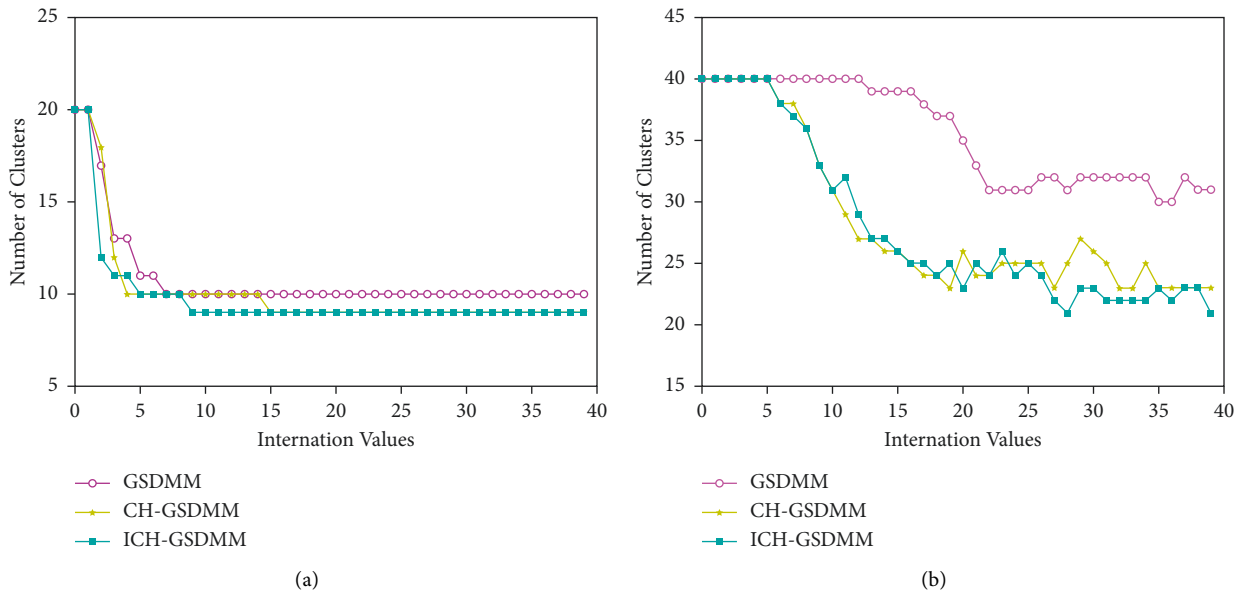

Figure 3: Number of clusters under various methods when iterations = 40. (a) Data set DS1. (b) Data set DS2.

(4) Effect Analysis of the Number of Clusters. To research the effect of the number of iterations on the number of clusters discovered by the ICH-GSDMM, CH-GSDMM, and GSDMM models, we set the initial cluster number parameter K of data set DS1 to 20, and the initial cluster number parameter K of data set DS2 to 40. Figure 3 displays the number of clusters discovered by ICH-GSDMM, CH-GSDMM, and GSDMM models at different iterations.

Figure 3(a) displays that the number of clusters discovered by the ICH-GSDMM, CH-GSDMM, and GSDMM models decreases rapidly and remains stable after approximately 9, 15, and 7 iterations, respectively. The closest order to the actual number of clusters is the ICH-GSDMM, CH-GSDMM, and GSDMM models.

Figure 3(b) shows that the number of clusters discovered by the ICH-GSDMM and CH-GSDMM models drops rapidly after about 6 iterations, while the GSDMM model drops rapidly after about 17 iterations, and the number of clusters finally discovered by the GSDMM model has the largest difference from the actual number of categories in the data set DS2. Both the ICH-GSDMM and CH-GSDMM models discover the number of clusters faster, and the ICH-GSDMM model found the number of clusters closest to the actual number of clusters after about 28 iterations. The number of documents in data set DS2 is 92.34% larger than that in data set DS1, which may be the reason why the number of clusters discovered by the ICH-GSDMM model in Figure 3(b) did not remain stable for a long time.

6. Conclusion

Compared with traditional topic modeling techniques, the GSDMM model is more suitable for short text clustering. However, in the GSDMM model, the Dirichlet prior distribution of words is supposed to be symmetric, i.e., all words are given the same prior distribution. When the model is constructed, all words are treated equally, which is obviously not realistic. To solve this problem, we proposed the ICH-GSDMM model. The improved chi-square statistics (ICH) method is the introduction of frequency, intraclass concentration, and interclass dispersion in the traditional chi-square statistical (CH) method. The ICH-GSDMM model is based on the ICH method to generate the Dirichlet prior distribution of important words of each category in the corpus to modify the traditional GSDMM model. Finally, we evaluate the internal and external clustering performance of traditional GSDMM, CH-GSDMM models, and the proposed ICH-GSDMM model in this paper. The results indicate that the internal evaluation index of the ICH-GSDMM model has improved greatly. The external evaluation index has improved except for NMI in the data set DS1. For the imbalanced data set DS1, the classification accuracy rate of minority classes is significantly improved, which also verifies the effectiveness of the model.

Future work will additional optimize the calculation method of the Dirichlet prior distribution of words in the GSDMM model and evaluate the impact of the number of important words in each category on the clustering effect to improve the ICH-GSDMM model and improve its external evaluation performance.

Data Availability

All data, models, and code generated or used during the study appear in the submitted article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China under Project No. 51967010.

References

[1] B. Ning, "A number of scientific and technical problems in intelligent transportation," *SCIENTIA SINICA Informationis*, vol. 48, no. 9, pp. 1264–1269, 2018.

[2] Q. Li, "Research on knowledge extraction method for high-speed railway signal equipment fault based on text," *Journal of the China Railway Society*, vol. 43, no. 3, pp. 92–100, 2021.

[3] Y. Zhao and X. Tian-Hua, "Text mining based fault diagnosis for vehicle on-board equipment of high speed railway signal system," *Journal of the China Railway Society*, vol. 37, no. 8, pp. 53–59, 2015.

[4] L. Yang, "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.

[5] Y. Wang, X. Zhang, and Q Song, "Characterization of the chromatin accessibility in an Alzheimer's disease (AD) mouse model," *Alzheimer's Research & Therapy*, vol. 12, no. 1, p. 29, 2020.

[6] Z. Zhong, T. Xu, W Feng, and T Tao, "Text case-based reasoning framework for fault diagnosis and predication by cloud computing," *Mathematical Problems in Engineering*, vol. 2018, no. 8, 10 pages, Article ID 9464971, 2018.

[7] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 49–58, 2017.

[8] R. Lu, "Character-level feature extraction method for railway text classification," *Computer Science*, vol. 48, no. 3, pp. 220–226, 2021.

[9] A. Blei, M. Ng, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[10] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 233–242, ACM, NY, USA, August 2014.

[11] J. A. Lossio-Ventura, S Gonzales, J Morzan, H Alatrista-Salas, T Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artificial Intelligence in Medicine*, vol. 117, no. 4, Article ID 102096, 2021.

[12] Y. Yang and O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420, DBLP, Pittsburgh,USA, July 1997.

[13] Y. Li, *Research on Word Segmentation and Feature Selection in Chinese Text Classification*, Vol. 7819, Jilin University, , Changchun, 2011.

[14] Z. Yan, "Improved method for text feature selection based on CHI," *Computer Engineering and Design*, vol. 37, no. 5, pp. 1391–1394, 2016.

[15] Y. Pei, "Study on improved CHI for feature selection in Chinese text categorization," *Computer Engineering and Applications*, vol. 47, no. 4, pp. 128–130, 2011.

[16] Z. Y. Xiong, "Improved approach to CHI in feature extraction," *Journal of Computer Applications*, vol. 28, no. 2, pp. 513-514, 2008.

[17] X. Lin, J. Lu, and L. Ran, "Automatic claon method of railway signal fault based on text mining," *Journal of Yunnan University (Natural Sciences Edition)*, vol. 44, no. 2, pp. 1–9, 2022.

[18] Q. Zhou and X. Li, "Research on short text classification method of railway signal equipment fault based on MCNN," *Journal of Railway Science and Engineering*, vol. 16, no. 11, pp. 2859–2865, 2019.

[19] G. Shang, "Research of fault feature extraction and diagnosis method for CTCS on-board equipment (OBE)based on labeled-LDA," *Journal of the China Railway Society*, vol. 41, no. 8, pp. 56–66, 2019.

[20] H. Jelodar, Y. Wang, and C Yuan, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.

[21] M. Pavlinek and V. Podgorelec, "Text classification method based on Self-Training and LDA topic models," *Expert Systems with Applications*, vol. 80, pp. 83–93, 2017.

[22] N. Agarwal, G. Sikka, and L. K. Awasthi, "Evaluation of web service clustering using Dirichlet Multinomial Mixture model

based approach for Dimensionality Reduction in service representation," *Information Processing & Management*, vol. 57, no. 4, Article ID 102238, 2020.

[23] B. Wang, "Service clustering based on GSDMM topic model,"vol. 1398, pp. 120–127, in *Proceedings of the 2021 International Conference on Applications and Techniques in Cyber Intelligence*, vol. 1398, Springer, Fuyang, China, June 2021.

[24] J. Yin and J. Wang, "A text clustering algorithm using an online clustering scheme for initialization," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'16*, pp. 1995–2004, New York, NY, USA, August 2016.

[25] R. Duan and C. Li, "An adaptive dirichlet multinomial mixture model for short text streaming clustering," in *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 49–55, IEEE, Santiago, Chile, December 2018.

[26] N. Peker and C. Kubat, "Application of Chi-square discretization algorithms to ensemble classification methods," *Expert Systems with Applications*, vol. 185, p. 49, Article ID 115540, 2021.

[27] P. Meesad, P. Boonrawd, and V. Nuipian, "A chi-square-test for word importance differentiation in text classification," in *Proceedings of the International Conference on Information and Electronics Engineering*, pp. 110–114, Text Classification, South Korea, Bangkok, Thailand, February 2011.

[28] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020.

[29] Q. Hu, J. Shenn, H. Jing, and W. Du, "Service clustering method based on description context feature words and improved GSDMM model," *Journal on Communications*, vol. 42, no. 8, pp. 176–187, 2021.

[30] A. Reddy, B. Tripathy, S. Nimje, S. Ganga, and K. Varnasree, "Performance analysis of clustering algorithm in data mining in R language," in *Proceedings of the International conference on soft computing systems*, pp. 364–372, Springer, Kerala, India, April 2018.

[31] W. Wang, B. Guo, Y. Shen, H. Yang, Y. Chen, and X. Suo, "Neural labeled LDA: a topic model for semi-supervised document classification," *Soft Computing*, vol. 25, no. 23, pp. 14561–14571, 2021.