IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 July 2023, accepted 21 August 2023, date of publication 24 August 2023, date of current version 31 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308520

RESEARCH ARTICLE

An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification

MINGMING ZHENG, KAIZHONG JIANG, RANHUI XU, AND LULU QI
School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China

Corresponding author: Kaizhong Jiang (kzjpub@sues.edu.cn)

This work was supported by the National Statistical Science Research Project of China under Grant 2020LY080.

ABSTRACT Nowadays, news text information is exploding, and people need more and more heterogeneous news content. Therefore, news text topic identification is needed to help viewers quickly and accurately screen and filter news related to their interests to save time and energy. The Latent Dirichlet Allocation(LDA) model is the most commonly used method for text topic identification. The optimal number of topics must be specified in advance when using the LDA model to extract topics in previous studies. However, selecting the too-large or the too-small number of topics significantly impacts the final results of LDA topic models, directly determining the quality of topic extraction. Moreover, the news text datasets from social media are very time-sensitive, and the combination of temporal and semantic modeling has not been considered in past studies of news topic identification. This paper proposes an adaptive optimal topic number determination method for fusing semantic and temporal information in news datasets to address the existing problems. Semantic and temporal are first extracted in this method as two different views. Then, density peak clustering of multi-view information fusion is performed based on the two obtained feature vectors. The clustering results are used as the final optimal number of topics. To demonstrate the effectiveness of the proposed method, this paper compares the performance of four traditional methods for determining the optimal number of topics with the performance of this paper's method on public datasets. The results show that the optimal number of topics considering semantic and temporal factors is significantly better than the other four traditional methods regarding F-value, PMI scores, and MI scores. It performs well in other indicators as well. The above experimental results show that the method proposed in this paper combines the temporal and semantic of news data to determine the optimal number of topics of news text, which can improve the accuracy of selecting the optimal number of topics in the LDA model and the effectiveness of the topic identification of news text to some extent. It can help viewers better understand and utilize the massive news text information. In addition, the method also broadens the idea of identifying and mining unique datasets from multiple perspectives.

INDEX TERMS LDA model, multi-view information fusion, optimal number of topics, news topics, social media data mining.

I. INTRODUCTION

In recent years, with the rapid development of Internet technology, the amount of information people obtain, especially text information, has grown exponentially. News information has become an essential source of information we receive daily because of its real-time, fast and widespread dissemination. Therefore, there is a need for methods such as news topic identification to enable viewers to quickly and accurately obtain helpful information from the vast amount of news datas. News topic identification refers to acquiring essential and representative information from large-scale news text datas and is widely used in topic detection and tracking [1], [2], [3]. Analyzing a large amount of news text data and mining hidden topic information can help users quickly obtain critical news topics and reduce the time spent browsing massive information. The analysis of hot spots and news information of interest to users can assist related fields

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

in predicting the future market and provide the necessary basis for judgment and decision making. It can also help the government monitor public opinion and prevent risks, all of which are of great practical significance.

The topic model is a statistical model that can discover abstract topics in text datasets and is often used to mine the semantic information implied in the text. It is a hot research subject in the field of natural language processing and a critical step in news topic identification technology. The topic model was first derived from Latent Semantic Indexing(LSI) [4], which is to transform word frequencies into a matrix of singular values by using singular value decomposition to remove the smaller singular value vectors and retain the largest few singular values. This approach can map document vectors and query vectors from word space to topic space but cannot handle synonyms and polysemous words. Hofmann [5] proposed the Probabilistic Latent Semantic Indexing (PLSI) model, representing the word process in the generated documents as a probabilistic statistical model. It is the second breakthrough of the topic model. However, the PLSI model only obtains a topic mixture ratio by fitting a finite number of documents in training set without describing the process of documents with a suitable probability distribution. This may lead to a linear increase of the model parameters with the number of texts. It isn't easy to obtain a suitable probability for documents outside the training set. Therefore, to address these issues, the Latent Dirichlet Allocation(LDA) topic model [6] is proposed.

The LDA topic model is modeled by introducing two Dirichlet prior distributions for the topics and the feature words corresponding to the topics, thus achieving a dimensionality reduction effect that allows the model parameters to not increase with increasing text datasets. Therefore, it is suitable for large-scale datasets. However, when modeling with LDA, the number of topics must be set in advance. The selection of the size of the number of topics will directly affect the model's performance. The current research still has no suitable method for determining the optimal number of topics. When using LDA models for news topic identification, many studies do not consider that news text data are very time-sensitive and do not fuse temporal and semantic modeling. People will focus on recent events rather than paying particular attention to past historical information. Based on the above problems, this paper adaptively determines the optimal number of topics for the LDA model in news topic identification by fusing semantic and temporal information. It makes the LDA model more effective in the identification of news topics. Moreover, it can help viewers to understand and utilize the massive news text information quickly and better.

II. RELATED WORK

A. DEVELOPMENT AND USEFULNESS OF THE LDA

LDA is a topic probability generation model that can represent essential text information in terms of the probability distribution of multiple latent topics. The result can be viewed as a Bayesian probabilistic graphical model containing three layers of documents, topics, and words. Blei et al. [6] solved the problem that the number of parameters increases as the text dataset increases in the probabilistic latent semantic analysis(PLSA) model by introducing the Dirichlet distribution. Thus, the LDA model is proposed. Since then, many scholars have improved the LDA model according to the actual situation. Rani and Lobiyal [7] proposed a tagged LDA extracted text summarization method based on topic modeling for Hindi novels and stories. They used a smoothing technique to process and diversify the content summaries. Then they evaluated the validity of the generated summaries based on gist diversity, retention ratio, and recall-oriented understudy for gisting evaluation(ROUGE) score. Ramage et al. [8] proposed a Labeled LDA model, which constrains the topics of LDA by defining the correspondence between potential topics of LDA and users' labels so that Labeled LDA can learn word-label correspondence directly and Labeled LDA possess more performance capabilities compared with traditional LDA. Liu et al. [9] proposed a multi-attribute LDA model, which considers microblogs' temporal and hashtag attributes in the topic analysis. With the time variable, the model can decide whether a word should appear in the trending topics or not, and with the hashtag variable, the model can rank the core words in front of the results, thus improving the model's expressiveness. When Watanabe and Baturo [10] investigated how the topic model automatically performs content analysis to make the content analysis more topic-specific and the topic classification more theoretically grounded, they proposed a Seeded Sequential LDA model, which was fitted using an iterative approach by ranging the number of topics, K, in the range of [3, 60] on the text of 1,000 speeches represented by the United Nations General Assembly. Moreover, the topic dispersion scores for three different topic granularities were calculated correspondingly. Finally, the topic dispersion scores determined K, and the optimal number of topics was utilized for relevant content analysis.

These improved LDA models have been widely used in natural languages processing and information retrieval, such as text dimensionality reduction, sentiment analysis, and topic identification. Crain et al. [11] used the LDA topic modeling approach to reduce dimensionality in text analysis with bag-of-words representations by collapsing terms with the same semantics together, thus identifying and eliminating terms with multiple meanings and obtaining low-dimensional document representations reflecting the concepts rather than the original ones. Wu et al. [12] proposed a short text clustering algorithm (Supervised Keyword Propagation for Latent Dirichlet Allocation, named SKP-LDA) based on sentiment word co-occurrence and knowledge symmetric feature extraction using the traditional LDA model. Since sentiment word co-occurrence entirely takes into account different short texts, by assigning sentiment polarity to short microblogging texts and by introducing the knowledge pairs of topic-specific and topic-relationship words into the LDA model, it can better and accurately respond to semantic information and effectively improve the reasonableness of the analysis of online public opinion. Bastani et al. [13], in order to identify themes in the text of Consumer Financial Protection Bureau (CFPB) complaint narratives and to explore the trends associated with them over time, they propose an analytical approach based on the LDA model, which incorporates temporal trends to assess the effectiveness of CFPB regulations and expectations for financial institutions in creating a consumer-oriented culture that allows for more efficient investigations to improve the experience of the consumer. Yang et al. [14], in order to mine Chinese government data governance policy themes and analyze the theme evolution path, using 443 government data governance policies above the provincial and ministerial levels in China as the data source for the experiment, LDA theme model identification was performed to identify the number of policy themes under different time windows. The peak of consistency scores was used as the optimal number of themes. The corresponding optimal number of themes for the three periods of germination, exploration, and development are finally identified. Zhou et al. [15] used Citibank's 2019 annual report as a data source and analyzed Environmental, Social, and Governance (ESG) factors with the LDA model. They calculated the complexity and drew complexity curves to determine the optimal number of themes, which provided a reference for analyzing economic, non-economic, Corporate Social Responsibility, and sustainable development.

B. OPTIMAL TOPIC NUMBER SELECTION IN LDA

Excellent results have been achieved in all of these areas. However, the selection of topic numbers significantly impacts the final result of the LDA topic model, which directly determines the quality of topic extraction. The current work is mainly based on the perplexity, hierarchical Dirichlet process, and Bayesian model to determine the size of the topic number, but all these methods have many shortcomings. The perplexity is the traditional method used in the LDA model proposed by Blei et al. [6]. They used an idea similar to exhaustive enumeration to take the number of topics used in the model with the lowest perplexity as the optimal number of topics. However, the number of topics selected by this method is usually large and has more disturbances. Teh et al. [16] proposed using a hierarchical Dirichlet process to adaptively obtain the optimal number, but this method is relatively inefficient. Later, some scholars proposed using the Bayesian nonparametric model to obtain the optimal number of topics, which has high computational complexity and does not have good generalization ability. In addition, in the text extraction results, each topic has a certain degree of similarity. The LDA model has the disadvantage that the extracted topics are not highly correlated. It cannot distinguish topics with higher similarity well, and it is also challenging to reflect the critical contents of the original text. To make the similarity between the topics extracted by the topic model as small as possible, many scholars have further improved the model from various perspectives, such as semantic, temporal, and corpus. Cao et al. [17] introduced the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to measure the relevance between topics and find the optimal LDA model by iteration. Li et al. [18] introduced category information into the existing LDA model feature selection algorithm and constructed a Support Vector Machine(SVM) multiclass classifier on the implied topic-text matrix to address the problem that traditional feature selection methods generally ignore the semantic relationships between words. Zhao et al. [19] proposed determining the optimal number of topics in the LDA model by the rate of change of the perplexity. He et al. [20] used JS scatter to calculate the relevance between topics and selected topics with high similarity. Huang et al. [21] introduced Term Frequency-Inverse Document Frequency (TF-IDF) based on the traditional LDA model, which can adjust the weight of words and perform fast calculations without considering the influence of word position in the document and can extract keywords in a shorter article length. Wang et al. [22] transformed the text topic cluster number selection problem into a clustering problem and demonstrated the effectiveness of clustering. Hasan et al. [23] proposed two new methods, normalized absolute coherence (NAC) and normalized absolute perplexity (NAP), for determining the optimal number of topics. In addition, Gan and Qi [24] constructed a comprehensive index of perplexity, isolation, stability, and coincidence to effectively identify the optimal number of topics in the LDA model based on the requirements for selecting the number of issues. Lu et al. [25] proposed a new adaptive LDA model, which employs new evaluation metrics to determine the optimal number of topics to be extracted from the SNS dataset, and finally verifies the model's performance through experiments. When Fernandes et al. [26] used the LDA model for clustering text, the model's output was probabilistic, which led to poor results in the first run. Therefore they used the LDA model recursively and utilized a hierarchical Dirichlet process to determine the selection of the optimal number of topics.These improved methods(such as only changing the data and considering a single perspective) refine topic extraction. In a short few scholars consider the method of analyzing from multiple perspectives together, and it has some shortcomings.

In summary, regardless of the traditional perplexity method, the hierarchical Dirichlet process, the Bayesian model, and the use of similarity to determine the optimal number of topics for LDA, there are still problems such as high computational complexity, high cost consumed, and poor topic extraction effect. Therefore, this paper uses text clustering to construct a new clustering method to determine the optimal number of topics for LDA.

C. NEWS TOPIC IDENTIFICATION

Although there are many news text sets, they all have the characteristics of incomplete sentences and vital timeliness, so unique methods are needed to identify news topics. In most related works, scholars are still based on the most classical LDA probabilistic topic model to identify news topics and obtain more accurate topic words by improving the model and adding some constraints [27]. Dai and Sun [28] performed news event identification by removing topic-specific stop words from each story and selecting some named entities as part of the feature. Trilling and Hoof [29] proposed the introduction of news events as a theoretical argument for the level of analysis and discuss several feasible computational methods to empirically detect news events in a large corpus of news reports in an unsupervised manner. Shao et al. [30] improved the traditional LDA model in two stages in order to address the shortcomings of text information omission and the slow speed of the algorithm when applied to news classification. Daud et al. [31] proposed a hyper-parameter optimized SVM to classify news articles into a category of articles with the specific same topic. Some scholars use the word to all documents and word co-occurrence for topic modeling to reduce the impact of sparse news text data.

Since news datasets are very time-sensitive compared to other datasets, many scholars have also added the time factor to the model and verified its effectiveness. Stilo and Velardi [32] disaggregated time series by using a symbolic aggregation algorithm and introduced them into news text clustering. Then, they verified through many experiments that time series have a particular influence. Some scholars have also modeled temporal dynamics by using the topic distribution of the previous time step in news text data to infer the current topic from continuous data. García-Méndez et al. [33] proposed an LDA model-based topic detection method for financial news, which helps investors to detect relevant financial events in unstructured text resources by considering relevance and temporality at the discourse level to identify relevant text in financial news and predictions in the text.

Through the analysis of related research works, this paper concluded that most news topic identification did not consider the LDA optimal number of topics, only added the temporal sequence to the model to adjust the output of the model, and did not model by fusion of the two views of the semantic information and the temporal information about the text itself. Therefore, this paper proposes a model that fuses semantic and temporal information. Then, multiview clustering is performed, and the number of clusters is set as the optimal number of topics for LDA in news topic identification.

III. METHODOLOGY

A. MODEL FRAMEWORK

This paper is roughly divided into three steps: data pre-processing for feature extraction, density peak clustering for multi-view information fusion and topic extraction performance comparison.

First, this paper pre-processes the obtained news text data. Due to the particular characteristics of the language itself, the pre-processing methods are different for different languages. Chinese requires word separation and removal of discontinued words. In contrast, English generally does not require word separation operations but pre-processing methods such as spell checking, stemming extraction, and word type reduction. However, to deal with multi-word phrases in English texts in the particular field of news, which will impact the effectiveness of the semantic feature extraction, the phrases in English have also been separated. Then, we take the word embedding for the pre-processed text to obtain the vector containing semantic information. Finally, the vector with temporal information is obtained by temporal analysis according to the time information corresponding to each document.

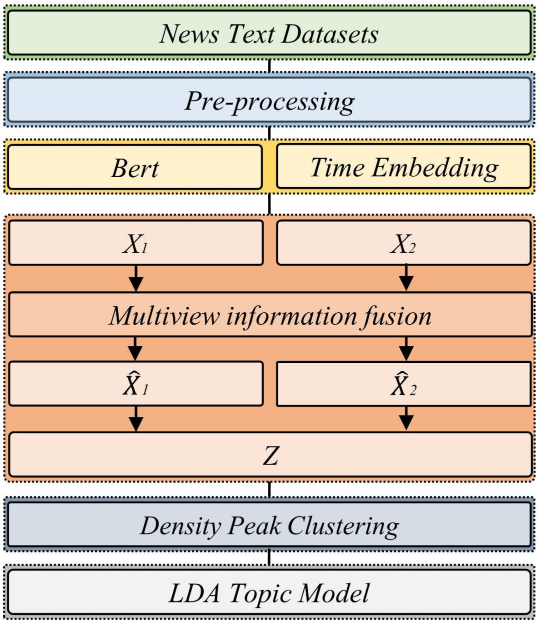


FIGURE 1. The framework of MVFDC.

This paper designs a density peak clustering method for multi-view information fusion using multi-level feature learning and comparison learning methods to train the processed semantic view vector and temporal view vector and obtain the feature vectors of two views fused. Then, the obtained feature vectors are subjected to density peak clustering to obtain the final clustering results [34], [35].

The clustering results are used as the number of topics for LDA for topic identification. This paper compare selecting the number of topics based on the perplexity MPI scores, C_UCI scores and UMass score by the experimental results. We have compared the advantages and disadvantages of topic extraction of the five methods.

The framework is shown in Fig. 1.

B. WORD EMBEDDING AND TEMPORAL VECTOR EXTRACTION

The traditional word embedding models are one-hot representation, bag-of-words model, n-gram model and other discrete representations, and distributed representations such as co-occurrence matrix. With the rapid development of neural networks, models such as Word Vector Representation(Word2Vec), Global Vectors for Word Representation(Glove), and Embeddings from Language Models (ElMO) have been proposed one after another. Transformer have been widely used recently due to the global modeling capability of the self-attention mechanism, which is effective in many tasks. BERT (Bidirectional Encoder Representations from Transformers) [36] is a pre-trained language representation model based on the Transformer model. Through unsupervised training, it learns context-sensitive word vector representations from large-scale text data and performs excellently on many natural language processing tasks. Due to its powerful representation capability and broad applicability, BERT has become one of the current state-of-the-art models. In the pre-training phase, BERT learns semantic representations through two tasks: language modeling and masked language modeling. The language modeling task requires BERT to predict the missing words in a given context, while the masked language modeling task requires BERT to predict the words in the masked input sentence parts. With this pre-training approach, BERT can model the contextual relationships of words and sentences and learn rich and accurate semantic representations. In the fine-tuning phase, the pre-trained BERT model is used for specific tasks, which involves adapting the model to specific downstream tasks such as sentiment classification and named entity identification. Many related studies have verified that BERT performs strongly in various natural language processing tasks.

So this paper uses the BERT pre-training model to obtain semantic word vectors, and is utilizing the masked language modeling. The Chinese language model has 12 layers of transformer blocks, 12 self-attention heads, 110 M parameters, and a large pre-training corpus of total 1B Chinese characters, including BookCorpus, Baidu encyclopedia, news corpus and other encyclopedias. The English language model has 24 layers of transformer blocks, 16 self-attention heads, 340 M parameters, and a large pre-training corpus of total 3.3 B words, including Wikipedia, BooksCorpus, other encyclopedias. The model also uses the technology of Whole Word Masking (WWM) [37] input for pre-training.

WWM is a masking technique whose goal is to better handle the masking task in participle languages such as Chinese to prevent words parsed into multiple subwords by the participle parser from being partially masked, thus improving the performance of the model. In the regular masking task, BERT randomly masks some words in the input sequence and then lets the model predict the masked words. Specifically, in the general model input sequence, each selected word has a 15% probability of being masked, of which 80% probability is replaced by a special masking symbol [MASK], a random word replaces 10% probability, and the remaining 10% probability is kept unchanged. There are these three types of masking in total. However, in participle languages such as Chinese, where words are composed of multiple characters and the participle is responsible for slicing the text into word units, masking subwords without masking the complete word may lead to confusion when the model handles such a task. The core idea of WWM is to mask a continuous sequence of words as a whole. Specifically, the original text is first divided into sequences of words using a disambiguator. Then for each word (such as phrases in Chinese) that is parsed by the disambiguator into multiple subwords, all the word subwords are masked. The advantage is that the model can consider the complete word context in the masking task and better understand and predict the masked words. It is important to note here that the WWM technique is needed when dealing with Chinese text because in Chinese, words usually consist of multiple characters and a lexer slices the text into word units; the motivation for WWM is to ensure that the model considers the whole word as a unit in the masking task, rather than dealing with each character or sub-word unit separately. By masking all the subwords of a word together, WWM allows the model to capture contextual information related to the complete word, thus improving the accuracy of the prediction.

In contrast, the WWM technique is unnecessary when dealing with English text because English words can usually be divided by spaces. In word-based or subword-based masking techniques (word-based Masked Language Model), masking individual characters or subwords can reasonably model the English vocabulary. However, many experiments prove that traditional individual masking degrades the model's performance when modeling in a specific domain or context due to many specific phrases in the English text. In contrast, whole-word masking can help the model to better capture the semantic and contextual information at the word level. In this paper, we need to obtain the contextual semantic information of specific news text data, so both the Chinese and English language models use the WWM techniques. We show a concrete sample operation of the WWM technique in Table 1. Only one of the three masking methods is shown in the table(substituting the special symbol [MASK]).

The Chinese model in this paper utilizes the Chinese open-source pre-training model jointly released by the Harbin Institute of Technology (HIT) and Xunfei Joint Laboratory. The model utilizes the Harbin Institute of Technology's self-developed Harbin Institute of Technology LTP (Language Technology Platform) word segmentation tool. A series of technical tools are used to reduce the cumulative error rate of word segmentation, such as lexicon matching, statistical language modeling, lexical annotation, unregistered word processing, and regularization rules. The English model utilizes the English open-source pre-trained model officially released by Google, which utilizes the WordPiece word segmentation tool. Some methods are also used to reduce the cumulative error rate of word segmentation, such as: designing word lists, handling unregistered words, considering contextual information, and word segmentation tagging. Finally, we input the pre-processed text data into the pre-trained model to obtain the semantic view vector for each word.

This paper implements temporal vector extraction using a similar signal construction in the Efficient Distributed Co-training over Wireless Networks(EDCoW) algorithm [38]. The obtained news dataset is divided according to days. There are T days of data in total, and the temporal vector of each word can be expressed in the following form.

$$S_w = [s_w(1), s_w(2), \cdots, s_w(T)] \tag{1}$$

where $s_w(t)$ at each sample point t is given by the fraction of DF-IDF, which is defined as:

$$s_w(t) = \frac{N_w(t) + \lambda_1}{N_t} * log \frac{\sum_{i=1}^{T} N_i}{\sum_{i=1}^{T} N_w(i) + \lambda_2} \tag{2}$$

where $N_w(t)$ denotes the number of news texts which contain word w and appear after sample point $t - 1$ but before t, and $N(t)$is the number of all news tests in the same period of time. The problem of sparse data and zero denominators is caused by certain words having zero occurrences in specific periods. Therefore, we use the Add-Delta smoothing method like λ_1 and λ_2 to solve the problem in this paper. Usually, λ_1 and λ_2 are taken as positive numbers less than one and are usually taken as 0.5. In this paper, we take λ_1, λ_2 in the range of [0.1,0.9] with a step size of 0.1, perform a simple pre-experiment on the two datasets we selected using feature learning and contrast learning, and use Accuracy and Purity to determine the optimal selection of λ_1, λ_2. The experimental results are shown in Table 2. Based on the results of the experiments, we set λ_1 to 0.5, λ_2 to 0.5 in processing Guardian data and set λ_1 to 0.3, λ_2 to 0.7 in processing Sogou Lab data in this paper. Finally, the temporal vector of each word can be obtained.

C. DENSITY PEAK CLUSTERING WITH MULTI-VIEW INFORMATION FUSION

In real life, an increasing amount of data are presented in the form of multiple views, and the analysis and study of multi-view data are beneficial for a more accurate understanding and judgment to some things. In this paper, we design a density peak clustering algorithm for multi-view information fusion (named MVFDC). To reduce the adverse effects generated by private information between views and avoid directly fusing the features of two views, we construct a multi-level feature learning model and use contrast learning to achieve the goal of consistency between different views. We use the trained model to obtain the information fusion of two views that are fused. Then, the feature vectors of the two views are concatenated. The obtained vectors are downscaled to perform density peak clustering. Finally, the decision map of density peak clustering is used to obtain the final number of clusters. The specific algorithm steps are as follows.

Step 1: Set initial parameters, input data from two views into the model, and map the original features into the low-level feature space using autoencoders.

Step 2: Transform the low-level features into high-level features and semantic labels using feature Multilayer Perceptron(MLP) and label MLP, respectively, to learn the semantics and clustering consistency of the two views in common.

Step 3: Modify the cluster labels obtained from the high-level features by the maximum matching formula and then fine-tuning the model by using the modified cluster labels.

Step 4: Output the fused feature vectors by the decoder, calculate the final total loss, and train the model by using the total loss and the small batch gradient descent algorithm.

Step 5: Use the trained model to perform forward inference to obtain the feature vectors of the standard semantics of the last two views after information fusion, concatenate the two views' vectors and cluster them using the density peak clustering algorithm, use the decision tree to determine the number of final clusters.

IV. EXPERIMENTS

A. DATA SOURCES AND PRE-PROCESSING

In this paper, we select two public datasets as shown in Table 3, one is the Guardian dataset of English news, and the other is the Sogou Lab dataset of Chinese news. The Guardian dataset has five categories: business, culture, politics, sports and technology, with a total of 51,283 articles, as well as the publication time corresponding to each news article. The Sogou Lab dataset has seven categories: sports, automobile, economy, entertainment, health, military, and education, with 700 articles and the corresponding news publication time. Ninety percent of the data are used as the training set to train the model, and ten percent are used as test sets to evaluate the model's strengths and weaknesses.

The text data of each news article are pre-processed. The English text needs to remove the symbols of non-English words and common stop words. In contrast, Chinese text needs to be divided into words in addition to removing non-Chinese characters and stop words due to the particular characteristics of Chinese text. Since this paper studies particular news texts, we have also divided English words based on particular English phrases in order to get better results.

B. IMPLEMENTATION

1) VIEW ACQUISITION

In this paper, the DF-IDF algorithm is used to extract the keywords of each category of articles. The first 3000 keywords of each category are selected for English, and the first 500 keywords of each category are selected for Chinese according to the data size. Then, the pre-trained language model BERT is used for word embedding to obtain the semantic vector. This paper select the pre-trained BERT model with a large corpus. The Whole Word Masking (wwm) training sample generation

TABLE 1. Sample generation of WWM technology.

Text Types	Clarification	Case
Chinese	Original Text	使用语言模型来预测下一个词的概率。
	Separate Text	使用 语言 模型 来 预测 下 一个 词 的 概率。
	Original Mask Input	使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 概 率。
	Full Word Mask Input	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK]。
English	Original Text	The soccer coach made strategic substitutions in the second half, aiming to inject fresh attacking impetus into the team.
	Separate Text	The soccer coach made strategic substitutions in the second half, aiming to inject fresh attacking impetus into the team.
	Original Mask Input	The soccer coach made [MASK] substitutions in the second half, aiming to inject fresh attacking [MASK] into the team.
	Full Word Mask Input	The soccer coach made [MASK] [MASK] in the second half, aiming to inject fresh [MASK] [MASK] into the team.

TABLE 2. Comparative experiments on the choice of parameters λ_1 and λ_2.

Parameters		Guardian		Sogou Lab	
λ_1	λ_2	ACC	PUR	ACC	PUR
0.1	0.9	0.4899	0.4899	0.3774	0.3951
0.2	0.8	0.4407	0.4603	0.3854	0.3920
0.3	0.7	0.4376	0.4379	**0.4317**	**0.4347**
0.4	0.6	0.4725	0.4790	0.4046	0.4046
0.5	0.5	**0.5129**	**0.5182**	0.3860	0.3829
0.6	0.4	0.4880	0.4932	0.4154	**0.4390**
0.7	0.3	0.4685	0.4725	0.3886	0.4009
0.8	0.2	0.4431	0.4480	0.3820	0.3934
0.9	0.1	0.4456	0.4456	0.3857	0.3977

TABLE 3. The information of the datasets in our experiments.

Datasets	#Language	#Samples	#Keywords	#Classes
Guardian	**English**	**700**	**3500**	**7**
Sogou Lab	**Chinese**	**51283**	**15000**	**5**

strategy is used. The details of the specific model parameters are shown in Table 4.

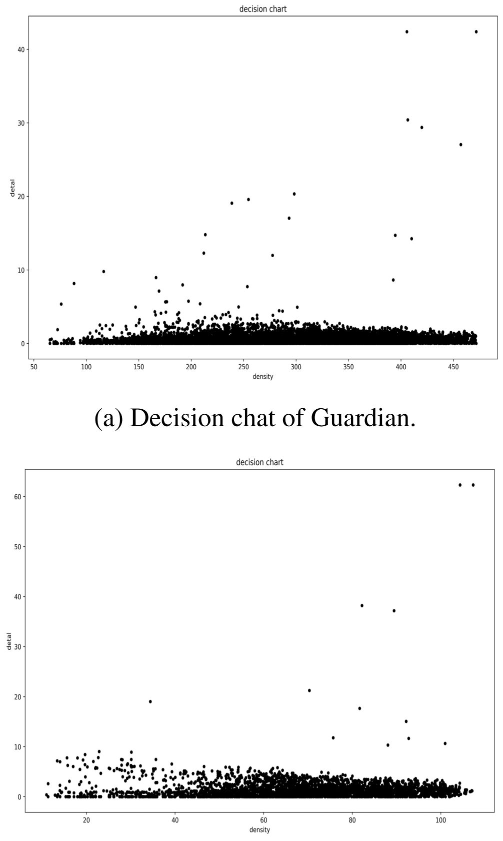
The selected keywords are used as keys, and the corresponding news text's publication time is used as the value to obtain a dictionary of words and time. Then, the number of news items published per day is used as the key to obtaining a dictionary of time and the number of texts. Finally, the temporal vector is obtained using the two formulas of Eq. (1) and Eq. (2).

2) OBTAIN THE VECTOR AFTER VIEW INFORMATION FUSION

This paper use the multi-level feature learning model and contrast learning to obtain the feature vectors after information fusion according to the two original view vectors. The corresponding vectors of the two views are concatenated together. To facilitate visual display, the TSNE package in Python was used for dimensionality reduction, and density peak clustering of the reduced data. Then we can determine the final clustering results according to the decision map.

V. RESULTS ANALYSIS

As shown below, the obtained word vectors and temporal vectors are used for density peak clustering for multi-view information fusion to obtain the final decision map.



(a) Decision chat of Guardian.

(b) Decision chat of Sogou Lab.

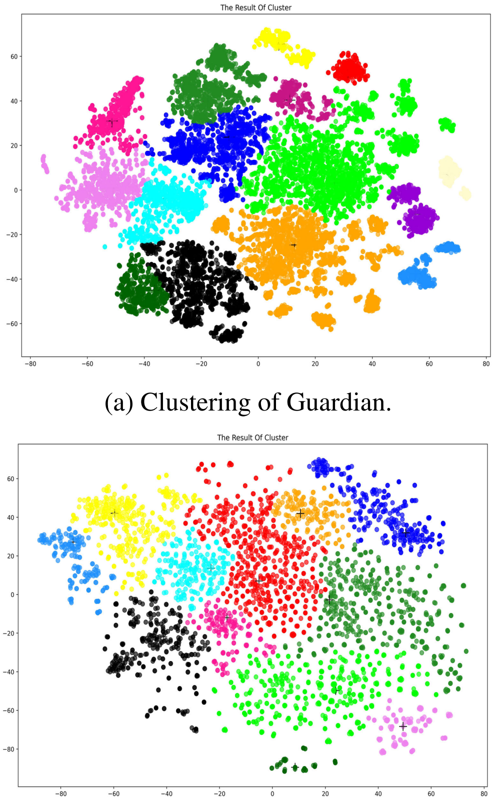
FIGURE 2. Decision chat for two datasets.

According to Fig. 2, there are 15 points with large horizontal and vertical coordinates in the Guardian dataset, which means that the data can be clustered into 15 classes. There are 12 points with large horizontal and vertical coordinates in the Sogou Lab dataset, which means that the data can be clustered into 12 classes. The visualization of the clustering results is shown in Fig. 3.

Evaluating the merit index of the LDA model generally uses the perplexity [6]. However, some work proves that using perplexity alone is not a good choice. Therefore, in order to make the experiments more adequate, this paper selects four baseline methods to evaluate the advantages and

TABLE 4. The information of the models in our experiments.

Datasets	Network Architecture	Masking	Type	Parameters
Guardian	12-layer, 768-hidden, 12-heads	WWM	large	110 M
Sogou Lab	24-layer, 1024-hidden, 16-heads	WWM	base	340 M



(a) Clustering of Guardian.

(b) Clustering of Sogou Lab.

FIGURE 3. Clustering chat for two datasets.

disadvantages of the LDA model as well as to select the optimal number of topics, which are Perplexity, Mean Pairwise Inter-topic Variability (MPI), C_UCI, and UMass [39]. Moreover, these four methods are compared with our proposed method.

The expression for the perplexity is defined as:

$$perplexity(D) = \left\{ \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{3}$$

where D denotes the test set in the corpus, M denotes a total of M documents, $N_d$denotes the number of words in each document d, W_d denotes the words occurring in document d, and $p(W_d)$ denotes the probability of the words W_d arising in the document.

First we define a formula for PMI calculation:

$$PMI(w_i, w_j) = \log \frac{Cofreq(w_i, w_j) + \epsilon}{freq(w_i) * freq(w_i)} \tag{4}$$

where $PMI(w_i, w_j)$ denotes the point mutual information for each pair of words (word i, word j), w_i and w_j denote word i and word j, $Cofreq(w_i, w_j)$ denotes the number of simultaneous occurrences of word i and word j within the co-occurrence window, $freq(w_i)$ and $freq(w_j)$ denote the number of occurrences of word i and word j in the document, and ϵ denotes for smoothing term.

The expression for the MPI is defined as:

$$MPI = \frac{\sum_{t} \sum_{w_i, w_j \in \ V} (PMI(w_i, w_j)/N_t)}{K} \tag{5}$$

The expression for the C_UCI is defined as:

$$C\_UCI = \frac{\sum_{t} \sum_{w_i, w_j \in \ V} PMI(w_i, w_j)/N_t(N_t - 1)/2}{K} \tag{6}$$

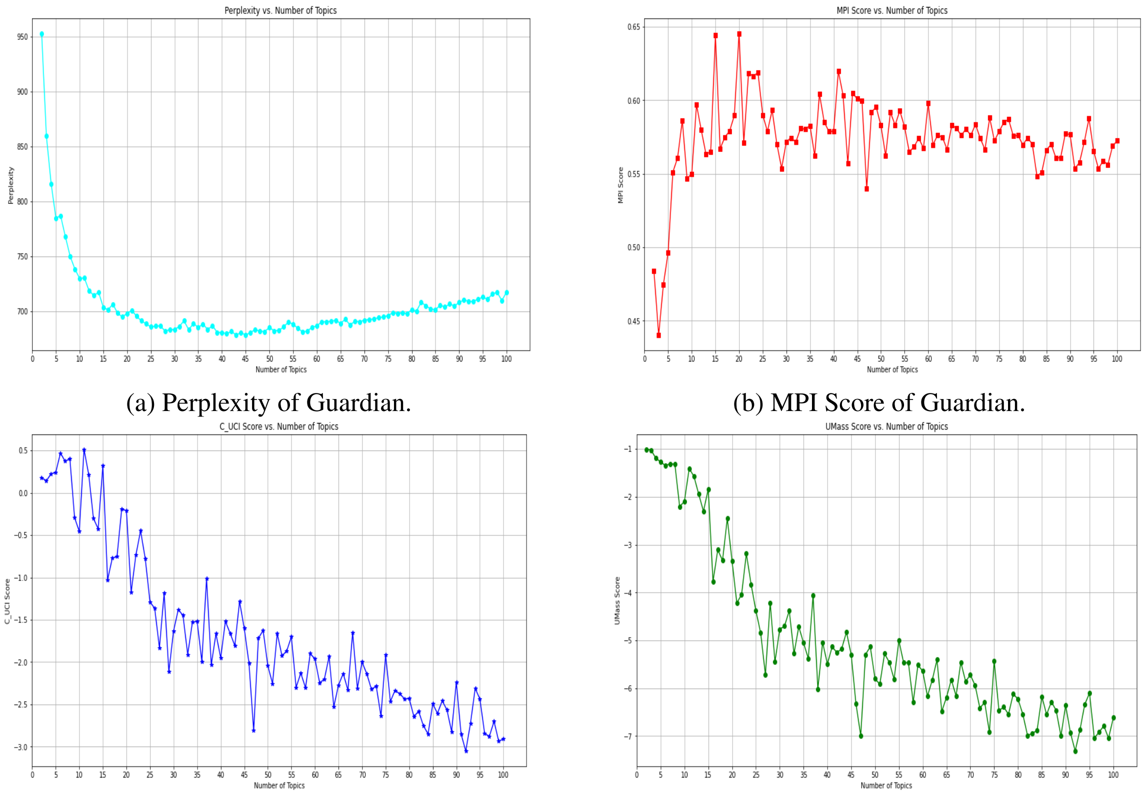
The expression for the UMass is defined as:

$$UMass = \frac{\sum_{t} \sum_{w_i, w_j \in \ V} PMI(w_i, w_j)}{K * N_t} \tag{7}$$

where t denotes topic t, N_t denotes the top N_t most representative words in topic t, and V denotes the set of these most representative words.

The higher the probability of the model on the test set, the smaller the perplexity, and the higher the MPI, the C_UCI score and the UMass score as the number of topics increases. In this paper, the number of topics is selected as [2:100] for experiments, the topic model is trained on the training set, the four indicators are calculated on the test set. The curves of four indicators under the different numbers of topics are plotted as follows.

From the results of Fig. 4 and Fig. 5, we can see that in the Guardian dataset, when the number of topics is 40, the perplexity level stops decreasing and tends to minimize, so the optimal number of topics can be considered 40. When the number of topics is 27, the MPI Score stops increasing, so the optimal number of topics can be considered 27. As the number of topics increases, U_uci Score and UMAsss Score are negative and getting smaller. Therefore we can determine the optimal number of topics as 11 and 8 based on these two methods. Similarly, in the Sogou Lab dataset, the method based on the perplexity can be considered 35, the method based on the MPI Score can be considered 45, the methods based on the U_uci Score and UMAsss Score can be



(a) Perplexity of Guardian.

(b) MPI Score of Guardian.

(c) C_UCI Score of Guardian.

(d) UMass Score of Guardian.

FIGURE 4. Curves of four indicators of Guardian dataset from 2 to 100.

considered 36 and 45 in the same way. Due to the small amount of data, the curves of the Sogou Lab dataset are more volatile. Therefore we choose to be the midpoint of the smooth fluctuation.

In this paper, the Precision, Recall F-value [40], AVGPMI(Average Point Mutual Information) and AVGMI (Average Mutual Information) are used to evaluate the extraction effect of the LDA model. The Precision is used to evaluate the proportion of correct topics among valid topics, the Recall is used to evaluate the proportion of correct topics extracted to the number of topics judged by experts, and the F-value is the summed average of the two. AVGPMI represents the average pointwise mutual information for all topics extracted, and AVGMI represents the average mutual information for all topics extracted. PMI can measure the degree of association between words and topics, and MI can measure the relevance of topics to each other.

$$P = \frac{T_{correct}}{T_{extract}}, R = \frac{T_{correct}}{T_{standard}}, F = \frac{2 * P * R}{P + R} \tag{8}$$

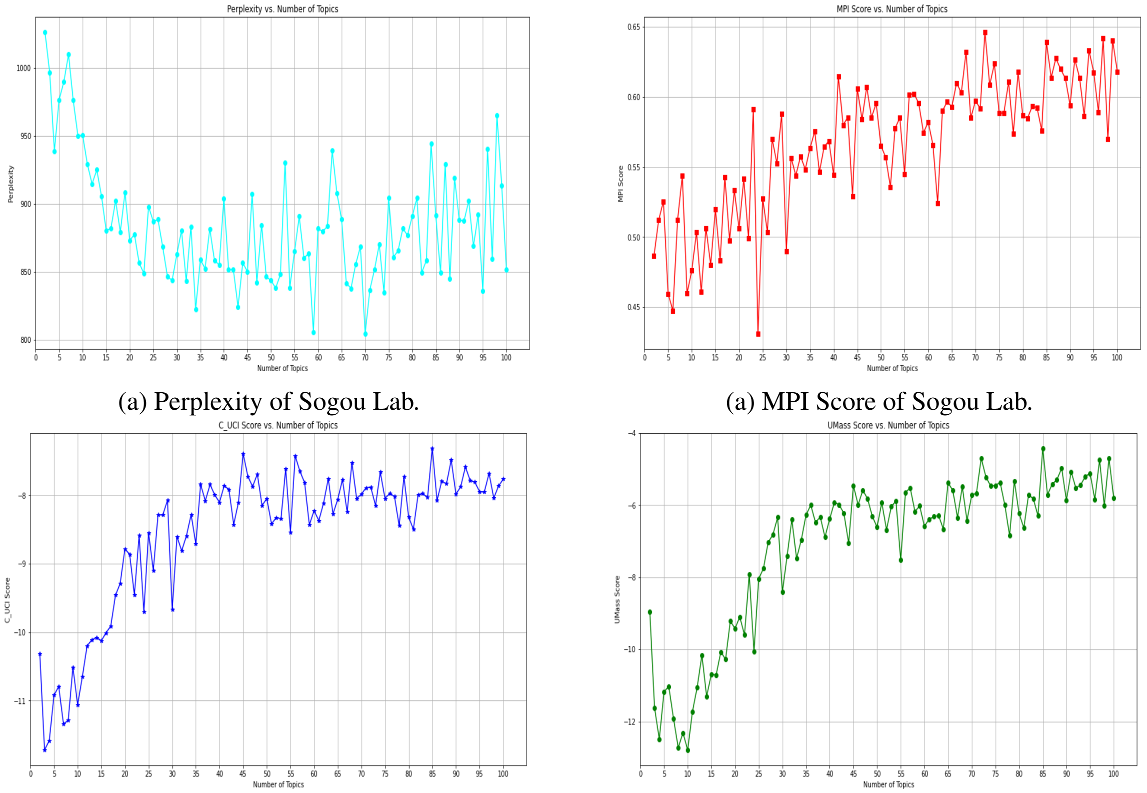
where $T_{extract}$ is the number of valid topics (the number of interfering items removed), $T_{standard}$ is the number of topics judged by the experts, and $T_{correct}$ is the number of correct topics (the number of valid topics included in the topics judged by the experts). The number of topics judged by the experts in this paper is the number of original categories in two datasets.

$$AVGPMI(t, w) = \frac{\sum_{t=1}^{K} log_2(P(t, w)/[P(t) * P(w)])}{K} \tag{9}$$

$$AVGMI(t_i, t_j) = \frac{\sum P(t_i, t_j) * log_2(P(t_i, t_j)/[P(t_i) * P(t_j)])}{K!} \tag{10}$$

where $P(t, w)$ denotes the joint probability of topic t and word w. $P(t)$ denotes the probability of topic t, and $P(w)$ denotes the probability of word w. $P(t_i, t_j)$ denotes the joint probability of topic t_i and topic t_j, $P(t_i)$ denotes the probability of topic t_i, and $P(t_j)$ denotes the probability of topic t_j. AVGPMI score is the higher indicates a stronger positive correlation between the words and topics, and the smaller AVGMI score indicates a strong differentiation before extracting the topics, and the better the LDA model extraction.

Table 5 shows the performance comparison of the two datasets under the five optimal topic number selection methods, from which it can be seen that some methods select a larger number of optimal topics. Therefore, the number of valid topics is both larger. The multi-view information fusion-based clustering method may not perform as well as the other traditional algorithms, but there are also many interference options. In terms of F-value PMI scores and MI scores, the multi-view information fusion-based clustering method



(a) Perplexity of Sogou Lab.

(a) MPI Score of Sogou Lab.

(c) C_UCI Score of Sogou Lab.

(d) UMass Score of Sogou Lab.

FIGURE 5. Curves of four indicators of Sogou Lab dataset from 2 to 100.

TABLE 5. The performance comparison of the five methods.

Datasets	Mothods	#	#	#	P	R	F	PMI	MI
Guardian	Perplexity	25	5	5	20.00%	100.00%	33.33%	10.9420	0.0477
	MPI	15	4	5	26.67%	80.00%	40.00%	10.8253	0.0492
	C_UCI	9	3	5	33.33%	60.00%	42.82%	10.4945	0.0489
	UMass	5	2	5	**40.00%**	40.00%	40.00%	10.1863	0.0539
	MVFDC	12	4	5	33.33%	**80.00%**	**47.06%**	**11.0212**	**0.0457**
Sogou Lab	Perplexity	13	5	7	38.46%	71.43%	50.00%	12.0840	0.0492
	MPI	13	6	7	46.15%	**85.71%**	59.99%	12.2114	0.0475
	C_UCI	12	6	7	50.00%	**85.71%**	63.16%	12.2298	0.0472
	UMass	13	5	7	38.46%	71.43%	50.00%	12.0840	0.0492
	MVFDC	7	5	7	**71.43%**	71.43%	**71.43%**	**12.3795**	**0.0409**

is significantly better than the other traditional algorithms. It also shows better results in Recall. Therefore, the new method extracted in this paper can improve the accuracy as well as the efficiency of news topic identification to a certain extent.

VI. CONCLUSION

With the massive popularity and high-speed development of Internet technology, news is rapidly and widely disseminated. The amount of news data is increasing, so the demand for mining and analyzing this kind of data is also increasing. This paper considers the characteristics of the data itself and combines the news content and temporal information for clustering to obtain the optimal number of topics for the LDA model. The experimental results show that the clustering method based on the fusion of semantic and temporal information to determine the optimal number of topics is significantly better than other traditional methods in terms of F-value, PMI scores MI scores, and it also expands the idea of analyzing and mining unique datasets from multiple perspectives.

However, modeling using this method may impact the results if an imbalance in the type of text data is encountered. Moreover, nowadays, there are many suitable methods to deal with the unbalanced classification problem, such as class-specific cost regulation extreme learning machine

(CCR-ELM) [41], so we consider how to improve further the performance of the method proposed in this paper in dealing with the unbalanced classification problem in our subsequent work. And the method requires textual datasets with temporal information, so the algorithm has some limitations on more general datasets. Currently, the combination of images, videos, and texts in multimedia has become a trend [42], and many news reports are accompanied by images and videos, which can also reflect the events more accurately. Therefore, in future work, we will focus on introducing images and videos into the model and try to build a more reasonable algorithm to analyze the data.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their constructive feedback on this manuscript.

REFERENCES

[1] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, ''Research on topic detection and tracking for online news texts,'' *IEEE Access*, vol. 7, pp. 58407–58418, 2019.

[2] J. Allan, ''Introduction to topic detection and tracking,'' in *Topic Detection and Tracking*. Berlin, Germany: Springer, 2002, pp. 1–16.

[3] L. Chen, H. Zhang, J. M. Jose, H. Yu, Y. Moshfeghi, and P. Triantafillou, ''Topic detection and tracking on heterogeneous information,'' *J. Intell. Inf. Syst.*, vol. 51, no. 1, pp. 115–137, Aug. 2018.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, ''Indexing by latent semantic analysis,'' *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[5] T. Hofmann, ''Probabilistic latent semantic indexing,'' in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, ''Latent Dirichlet allocation,'' *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[7] R. Rani and D. K. Lobiyal, ''An extractive text summarization approach using tagged-LDA based topic modeling,'' *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 3275–3305, Jan. 2021.

[8] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, ''Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 248–256.

[9] G. Liu, X. Xu, Y. Zhu, and L. Li, ''An improved latent Dirichlet allocation model for hot topic extraction,'' in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput.*, Dec. 2014, pp. 470–476.

[10] K. Watanabe and A. Baturo, ''Seeded sequential LDA: A semi-supervised algorithm for topic-specific analysis of sentences,'' *Social Sci. Comput. Rev.*, pp. 1–25, May 2023.

[11] S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha, ''Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond,'' in *Mining Text Data*. Berlin, Germany: Springer, 2012, pp. 129–161.

[12] D. Wu, R. Yang, and C. Shen, ''Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm,'' *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 1–23, Feb. 2021.

[13] K. Bastani, H. Namavari, and J. Shaffer, ''Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints,'' *Exp. Syst. Appl.*, vol. 127, pp. 256–271, Aug. 2019.

[14] Q. Yang, ''LDA-based topic mining research on China's government data governance policy,'' *Social Secur. Admin. Manag.*, vol. 3, no. 2, pp. 33–42, 2022.

[15] Z. Zhou, M. Liu, and Z. Tao, ''Quantitative analysis of Citi's ESG reporting: LDA and TF-IDF approaches,'' *Financial Eng. Risk Manag.*, vol. 6, no. 3, pp. 53–63, 2023.

[16] Y. Teh, M. Jordan, M. Beal, and D. Blei, ''Sharing clusters among related groups: Hierarchical Dirichlet processes,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 1–8.

[17] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, ''A density-based method for adaptive LDA model selection,'' *Neurocomputing*, vol. 72, nos. 7–9, pp. 1775–1781, Mar. 2009.

[18] K. Li, J. Xie, X. Sun, Y. Ma, and H. Bai, ''Multi-class text categorization based on LDA and SVM,'' *Proc. Eng.*, vol. 15, pp. 1963–1967, Jan. 2011.

[19] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, ''A heuristic approach to determine an appropriate number of topics in topic modeling,'' in *BMC Bioinformatics*. vol. 16. Berlin, Germany: Springer, 2015, pp. 1–10.

[20] J. He, X. Chen, M. Du, and H. Jiang, ''Topic evolution analysis based on improved online LDA model,'' *J. Cent South Univ.*, vol. 46, no. 2, pp. 547–553, 2015.

[21] L. Huang, J. Ma, and C. Chen, ''Topic detection from microblogs using T-LDA and perplexity,'' in *Proc. 24th Asia–Pacific Softw. Eng. Conf. Workshops (APSECW)*, Dec. 2017, pp. 71–77.

[22] H. Wang, J. Wang, Y. Zhang, M. Wang, and C. Mao, ''Optimization of topic recognition model for news texts based on LDA,'' *J. Digit. Inf. Manag.*, vol. 17, no. 5, p. 257, Oct. 2019.

[23] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, and M. J. Islam, ''Normalized approach to find optimal number of topics in latent Dirichlet allocation (LDA),'' in *Proc. Int. Conf. Trends Comput. Cognit. Eng.* Cham, Switzerland: Springer, 2021, pp. 341–354.

[24] J. Gan and Y. Qi, ''Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example,'' *Entropy*, vol. 23, no. 10, p. 1301, Oct. 2021.

[25] F. Lu, B. Shen, J. Lin, and H. Zhang, ''A method of SNS topic models extraction based on self-adaptively LDA modeling,'' in *Proc. 3rd Int. Conf. Intell. Syst. Design Eng. Appl.*, Jan. 2013, pp. 112–115.

[26] N. Fernandes, A. Gkolia, N. Pizzo, J. Davenport, and A. Nair, ''Unification of HDP and LDA models for optimal topic clustering of subject specific question banks,'' 2020, *arXiv:2011.01035*.

[27] D. Walter and Y. Ophir, ''News frame analysis: An inductive mixed-method computational approach,'' *Commun. Methods Measures*, vol. 13, no. 4, pp. 248–266, Oct. 2019.

[28] X. Dai and Y. Sun, ''Event identification within news topics,'' in *Proc. Int. Conf. Intell. Comput. Integr. Syst.*, Oct. 2010, pp. 498–502.

[29] D. Trilling and M. van Hoof, ''Between article and topic: News events as level of analysis and their computational identification,'' *Digit. Journalism*, vol. 8, no. 10, pp. 1317–1337, Nov. 2020.

[30] D. Shao, C. Li, C. Huang, Y. Xiang, and Z. Yu, ''A news classification applied with new text representation based on the improved LDA,'' *Multimedia Tools Appl.*, vol. 81, no. 15, pp. 21521–21545, Jun. 2022.

[31] S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar, ''Topic classification of online news articles using optimized machine learning models,'' *Computers*, vol. 12, no. 1, p. 16, Jan. 2023.

[32] G. Stilo and P. Velardi, ''Efficient temporal mining of micro-blog texts and its application to event discovery,'' *Data Mining Knowl. Discovery*, vol. 30, no. 2, pp. 372–402, Mar. 2016.

[33] S. Garcia-Mendez, F. de Arriba-Pérez, A. Barros-Vila, F. J. Gonzàlez-Castano, and E. Costa-Montenegro, ''Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with latent Dirichlet allocation,'' *Appl. Intell.*, vol. 53, pp. 1–19, Mar. 2023.

[34] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, ''Multi-level feature learning for contrastive multi-view clustering,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 16051–16060.

[35] A. Rodriguez and A. Laio, ''Clustering by fast search and find of density peaks,'' *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.

[37] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, ''Pre-training with whole word masking for Chinese BERT,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021.

[38] J. Weng and B.-S. Lee, "Event detection in Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, 2011, vol. 5, no. 1, pp. 401–408.

[39] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 100–108.

[40] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proc. 10th Annu. Joint Conf. Digit. Libraries*, Jun. 2010, pp. 215–224.

[41] W. Xiao, J. Zhang, Y. Li, S. Zhang, and W. Yang, "Class-specific cost regulation extreme learning machine for imbalanced classification," *Neurocomputing*, vol. 261, pp. 70–82, Oct. 2017.

[42] P. Meel and D. K. Vishwakarma, "Multi-modal fusion using fine-tuned self-attention and transfer learning for veracity analysis of web information," *Exp. Syst. Appl.*, vol. 229, Nov. 2023, Art. no. 120537.

MINGMING ZHENG was born in Mianyang, Sichuan, China, in 1998. He received the bachelor's degree in applied statistics from the Chengdu College of Arts and Sciences, Chengdu, China, in 2021. He is currently pursuing the master's degree in statistics with the Shanghai University of Engineering Science, Shanghai, China. His research interests include data mining, machine learning, deep learning, and natural language processing.

KAIZHONG JIANG received the Ph.D. degree in systems analysis and integration from East China Normal University, Shanghai, China, in 2008. He is currently an Associate Professor with the School of Mathematics, Science and Statistics, Shanghai University of Engineering Science. He has published more than 30 journal articles. His research interests include business WEB data mining, machine learning, information retrieval and knowledge retrieval, complex networks and complex systems, and algorithm research and design.

RANHUI XU was born in Dongying, Shandong, China, in 1999. He received the bachelor's degree in statistics from Jiaxing University, Jiaxing, China, in 2021. He is currently pursuing the master's degree in statistics with the Shanghai University of Engineering Science, Shanghai, China. His research interests include data mining, machine learning, deep learning, and applications in computer vision.

LULU QI was born in Dongying, Shandong, China, in 1998. She received the bachelor's degree in economic statistics from Qufu Normal University, Qufu, China, in 2020. She is currently pursuing the master's degree in statistics with the Shanghai University of Engineering Science, Shanghai, China. Her research interests include machine learning and few-shot image classification.