Text Information Data Mining Method in Natural Language Processing Tasks

Shengguo Guo*, Dandan Xing
School of Information Engineering, Zhenzhou College of Finance and Economics, Zhengzhou, China

***Abstract*—Text mining methods often rely on a single data source or simple word frequency statistics, making it difficult to capture multi-source text semantic associations and local contextual dependencies, resulting in poor mining accuracy. Therefore, a method for text information data mining in natural language processing tasks is proposed. Using Python web crawlers to obtain multi-source text data, after preprocessing such as cleaning, segmentation, and removal of stop words, a Vector Space Model (VSM) is used for text representation, and a TF-IDF (Term Frequency Across Document Frequency) weight optimization mechanism is introduced to enhance feature semantic representation. On this basis, a semantic enhancement system is constructed based on the BERT classification model in the field of natural language processing. Through the self-attention mechanism of multi-layer Transformer encoders, semantics are aggregated to effectively capture local contextual dependencies, and context-sensitive word vectors are generated by the output layer. Finally, by fine-tuning the parameters of the BERT (Bidirectional Encoder Representation from Transformers) model and combining it with the Softmax function, precise mining of text information data categories was achieved. The experimental results show that in the embedding experiment of sports news headlines, this method can form a semantic aggregation structure with clear domain logic for word vectors; In the cross domain short text classification experiment, the overall accuracy of this method on the dataset reached 95.7%, which was 19.5% and 18.7% higher than the comparative methods, effectively solving the cross domain ambiguity problem in natural language processing.**

Keywords*—Natural language processing; text information; data mining; VSM; TF-IDF; BERT

I. Introduction

Text information data is widely distributed in scenarios such as social media, academic platforms, enterprise systems, and medical databases, covering diverse types such as online reviews, academic papers, financial annual reports, and electronic medical records. These data continue to grow at a rate of hundreds of millions per day, forming a huge unstructured information cluster. Different from structured data, text information data exhibits significant looseness and complexity: online jargon, abbreviations, and emojis are mixed in social media comments; academic literature involves professional terms and complex sentence patterns; industry jargon and ambiguous expressions exist in enterprise reports; medical records contain medical codes and colloquial records. These characteristics make it difficult for computers to directly and efficiently process and analyze text information [1]. Facing such complex and disorderly text information data, how to extract valuable knowledge from it has become an important topic in the field of data science. Studying text information data mining methods aims to break through the unstructured barriers of text and extract implicit patterns, trends, and relationships from massive texts [2]. Taking an e-commerce platform as an example, by mining consumer reviews, not only can specific pain points of products in dimensions such as function, quality, and service be identified, but also the evolution trend of user needs over time can be traced; in the field of academic research, in-depth mining of literature can reveal interdisciplinary hotspots, technology evolution paths, and key unsolved problems; in the medical field, analysis of medical record texts can discover potential associations between disease symptoms and treatment plans, and even predict the incidence patterns of diseases [3]. Through these mining efforts, disorderly text data is transformed into structured knowledge, achieving a qualitative change from "data accumulation" to "knowledge precipitation", providing a core basis for decision-making, technological innovation, and problem-solving.

In the research on text data mining in the industry, Kim, H. J. et al. proposed a conditional generative adversarial multiple imputation network based on unsupervised data mining. An auxiliary classifier was developed through a pre-training algorithm combined with implicit class labels. The sorting point recognition clustering structure algorithm of fuzzy clustering was used to learn implicit classification information, and multiple imputation was applied to the original energy dataset to improve the robustness and reliability of the imputation results [4]. However, when explicit class labels are not available, it is difficult to effectively capture the semantic associations of multi-source texts, resulting in poor accuracy of mining results. Rahman, U. et al. used two classification algorithms, Naive Bayes and Support Vector Machine, to perform binary classification on the text data generated by the regular maintenance of industrial plant equipment. The model was trained and tested by analyzing the labeled data to achieve the early classification of equipment failures, thereby improving the industry maintenance plan [5]. However, this method does not consider capturing the long-distance dependence relationships between different words in the text, which may lead to a decrease in the classification mining accuracy in complex fault scenarios. Kumar, M. R. P. et al. mined context-related subjective words in the text through a context-based sentiment dictionary, calculated the weight scores of subjective terms and complete reviews, and then classified the text sentiment with the help of a bidirectional LSTM model. At the same time, an oversampling technique was used to generate new synthetic text samples using semantic information to solve the

* Corresponding author.

data imbalance problem and achieve sentiment classification mining of imbalanced corpora [6]. However, the context-based sentiment dictionary needs to be manually defined or pre-annotated, and it is difficult to cover new sentiment expressions in dynamic contexts, resulting in poor capture of the semantic associations of multi-source texts and poor adaptability to out-of-domain texts. Kim, H. et al. proposed a flexible periodic pattern data mining method for incremental time series databases, which processes data streams in real time by enhancing the data structure and mines periodic patterns containing uncertain events, providing decision support for scenarios such as oil price fluctuation prediction and traffic congestion analysis [7]. However, this method cannot accurately express the semantic differences of words in different contexts, easily leading to insufficient accuracy of text representation and poor mining performance.

In summary, existing text information data mining methods still have significant limitations in deep semantic modeling and cross-domain generalization ability: most models fail to effectively integrate the statistical features of text with global contextual semantics, resulting in insufficient semantic understanding of polysemous words, domain-specific terminology, and cross-domain short texts, thereby restricting further improvement in mining accuracy and practicality. Natural Language Processing (NLP), as the core technology for processing text data, provides crucial support for text information data mining [8]. In recent years, NLP technologies driven by deep learning, such as pre-trained language models like BERT and GPT, have significantly improved the ability to understand text semantics through learning from massive corpora, greatly promoting the development of text information data mining [9]. These technologies can accurately capture the sentiment tendency, entity relationships, and topic context in text, demonstrating great value in practical applications: in the e-commerce field, mining the semantic features of consumer reviews can assist in product optimization; in the medical industry, mining medical record texts helps with clinical decision-making; in academic research, literature semantic analysis can help keep up with the cutting-edge of the discipline [10]. Therefore, in view of the limitations of the above methods, this study focuses on the research of text information data mining methods in natural language processing tasks, in order to promote the cross-innovation of natural language processing and data mining technologies, and provide efficient transformation solutions from text data to structured knowledge for various industries. The specific technical route is as follows:

- Firstly, a Python web crawler is used to construct a multi-source heterogeneous text corpus, and strict cleaning, segmentation, and stop word preprocessing are performed to provide a high-quality, cross-domain training and evaluation foundation for the model, overcoming the bias problem caused by a single data source.
- Then, based on the vector space model representation, the TF-IDF weight optimization mechanism is introduced. This step goes beyond simple word frequency statistics and can enhance key semantic features with high discrimination, providing strong feature inputs rich in statistical information for subsequent deep models and improving the discriminative power of semantic representation from the perspective of feature engineering.
- Subsequently, a semantic enhancement system based on the BERT classification model was constructed, which utilizes the self-attention mechanism of the Transformer encoder to dynamically aggregate global contextual information, accurately capturing long-distance dependencies and polysemy between words, fundamentally solving the core deficiency of weak contextual modeling ability in traditional models.
- Finally, the optimized TF-IDF features are deeply fused with the context vector of BERT through fine-tuning strategies, and precise classification is achieved using the Softmax function. This fusion scheme not only inherits the advantages of traditional methods with clear features, but also has the semantic understanding ability of deep models.
- Design a cross-domain short text classification experiment to verify the comprehensive performance of the proposed framework in handling polysemous words and domain-specific terms from dimensions such as semantic representation quality, computational efficiency, and term robustness on a multi-domain news headline dataset, including sports, entertainment, and real estate.

II. Optimize Text Feature Vectors Based on TF-IDF

In the text information data mining task in the field of Natural Language Processing (NLP), the core is to convert unstructured text into computable semantic information, and data acquisition and preprocessing are the basic links to achieve this goal [11]. This study applies web crawler technology based on Python to crawl the required text data from various text data sources, such as social media, news texts, and academic literature. Collect information such as text content and publication time, and store the data in text format to provide the original corpus for subsequent text information data mining [12].

A. Text Information Data Preprocessing

The preprocessing process of text information data is shown in Fig. 1, and the specific process is as follows:

1) Text data cleaning: Remove the possible special symbols and extra blanks in the crawled text data. Extra blanks may separate complete words, resulting in incorrect word segmentation, reducing the word segmentation accuracy, and thus affecting the final text information data mining effect. Therefore, it is necessary to ensure the standardization of text data through cleaning [13].

2) Chinese text word segmentation: Select the jieba word segmentation library based on the Chinese word library in Python to perform word segmentation on the text data. Split the text content into characters, words, or phrases so that the computer can accurately recognize the words in the text. This

method has high word segmentation accuracy and fast speed, which can improve the accuracy of subsequent text classification, clustering, and other tasks [14].

3) Remove obsolete terms: Compare with the obsolete terms library to filter the text data, delete words such as modal particles, prepositions, and pronouns that have no actual meaning and may affect the text mining results, increase the keyword density of the text, and at the same time reduce the feature dimension, thereby improving the accuracy and efficiency of text mining [15].

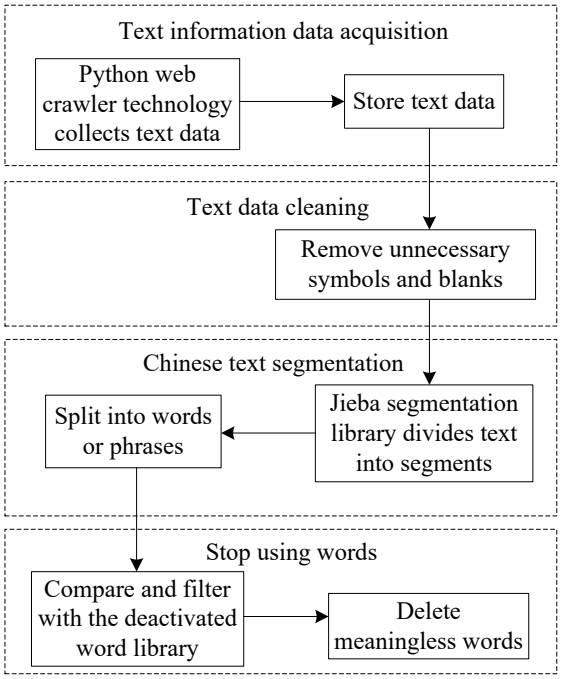

Fig. 1. Preprocessing process of text information data.

B. Text Representation Based on Vector Space Model

Text representation, as a key bridge connecting natural language and machine understanding, directly determines the accuracy and efficiency of subsequent mining tasks. After acquiring and preprocessing text data, it is necessary to further transform unstructured text into a structured representation that can be processed by a computer, laying a foundation for in-depth semantic analysis. There are mainly three common text representation models: the Boolean model, the probabilistic model, and the vector space model. The vector space model (VSM), as one of the most widely used and effective methods currently, has its technical advantages directly reflected in the processing efficiency and mining accuracy of actual text data [16]. The smallest data unit in the vector space model is the feature term, and words, phrases, and word groups can all be used as feature terms for processing. Regarding the text c as a n -dimensional vector in the vector space, as shown in Eq. (1):

$$c=\left(\left(\varepsilon_1,\omega_1\right),\left(\varepsilon_2,\omega_2\right),\cdots,\left(\varepsilon_n,\omega_n\right)\right) \tag{1}$$

where, ε_i represents the i -th feature term of the text c; ω_i represents the feature weight corresponding to the i -th feature term of the text c, and the magnitude of the feature weight indicates the amount of text category information contained in this feature.

C. Weight Optimization Based on TF-IDF

The text data representation based on the vector space model realizes the structured processing of text through the vector mapping of feature terms and weights. However, this method relies on word frequency statistics and lacks semantic association modeling. Therefore, the TF-IDF (Term Frequency-Inverse Document Frequency) weight optimization representation method is introduced to enhance the semantic representation ability on the basis of retaining the VSM feature weight system and adapting to the complex semantic requirements in text information data mining [17]. For example, in VSM, "computer" and "PC" are regarded as independent dimensions with no associated weights, while TF-IDF ensures that both obtain high weights in scientific and technological texts through IDF, indirectly improving the similarity of semantically related texts.

For the text collection $C=\left\{c_1,c_2,\cdots,c_M\right\}$, M is the total number of texts in the text collection C. Following the feature term processing logic of VSM, the feature term set $\left\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\right\}$ is obtained. Taking each feature term ε_i as a dimension of the vector space, the weight ω_i of each feature term ε_i in the document is calculated through TF-IDF, as shown in Eq. (2) to Eq. (4):

$$\omega_i=\omega\left(\varepsilon_i,C_j\right)=\frac{TF\left(\varepsilon_i,C_j\right)\times IDF\left(\varepsilon_i\right)}{\sqrt{\sum_{\varepsilon_i\in C_j}\left[TF\left(\varepsilon_i,C_j\right)\times IDF\left(\varepsilon_i\right)\right]^2}} \tag{2}$$

$$TF\left(\varepsilon_i,C_j\right)=\frac{D\left(\varepsilon_i,C_j\right)}{D\left(C_j\right)} \tag{3}$$

$$IDF\left(\varepsilon_i\right)=\log\left(\frac{M}{D\left(\varepsilon_i\right)}+0.01\right) \tag{4}$$

where, $TF\left(\varepsilon_i,C_j\right)$ represents the relative frequency of the feature term ε_i in the single text C_j; $IDF\left(\varepsilon_i\right)$ is the inverse document frequency, which measures the discrimination ability of the feature term ε_i in the entire text collection. If the feature term ε_i appears in more texts ($D\left(\varepsilon_i\right)$ is larger), then its discrimination degree is lower. On the contrary, if it only appears in a few texts, the higher the IDF value, the greater the weight; $D\left(\varepsilon_i,C_j\right)$ is the absolute number of occurrences of the feature term ε_i in the text C_j; $D\left(C_j\right)$ is the total

number of feature terms in the text C_j; $D(\varepsilon_i)$ is the number of texts containing ε_i.

Assume that the text $C_j = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_t)$ contains t feature terms. Multiply the weight of each feature term to the corresponding vector dimension to obtain the weighted text feature vector y_i, as shown in Eq. (5):

$$y_i = \varepsilon_i \times \omega(\varepsilon_i, C_j) \tag{5}$$

The feature vectors of each text in the text collection C is shown in Eq. (6):

$$\phi_i = \frac{\sum_{j=1}^{t} y_j}{|C_i|} \tag{6}$$

where, $\sum_{j=1}^{t} y_j$ is the semantic contribution that integrates all features within the text. For example, in scientific and technological texts, the vector weights of words such as "artificial intelligence" and "big data" will be accumulated. $|C_i|$ is the text length normalization factor.

III. Text Information Data Mining Based on Semantic Enhancement of BERT Classification Model

After optimizing the text feature vector based on TF-IDF, the BERT classification model (Bidirectional Encoder Representations from Transformers) is further introduced to achieve the upgrade from statistical representation to semantic representation, thereby capturing the long-distance dependency relationships between different words in the text and automatically learning the importance weights between words, so as to generate more accurate and context-sensitive word vectors [18]. This model adopts a bidirectional variant of the encoder-decoder architecture and consists of an input layer, an encoding layer, and an output layer. The model structure is shown in Fig. 2. The design of this model structure directly serves the classification task in text information data mining:

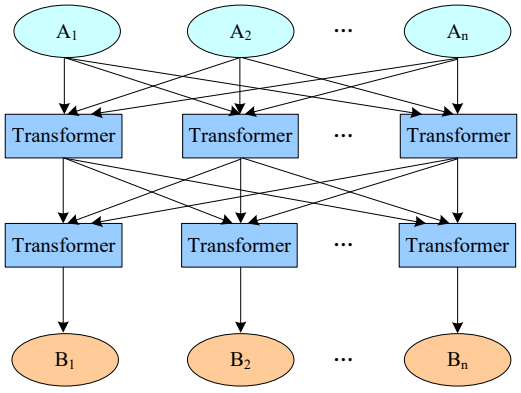

Fig. 2. Structure of BERT classification model.

The input layer combines the TF-IDF text feature vector representation $\{\phi_1, \phi_2, \cdots, \phi_n\}$ with the original text information to generate a composite text input sequence $\{A | A_1, A_2, \cdots, A_n\}$; the encoding layer is stacked by multiple layers of Transformer encoders, and each layer realizes semantic aggregation through the self-attention mechanism and the feed-forward neural network; the output layer $\{B | B_1, B_2, \cdots, B_n\}$ represents the generated word vector. A_i corresponds to B_i, and i is the i -th word of the input text.

Under the self-attention mechanism, the input vector is characterized by three vectors: the query vector Q, the key vector K, and the value vector V. The semantic association weights between words are dynamically calculated through triple operations [19], and the context of the TF-IDF feature vector representation is corrected. For example, "bank" generates different encodings in the contexts of "financial bank" and "riverbank". Input vector calculation expressions are shown in Eq. (7) to Eq. (9):

$$Q = f(q(A)) \tag{7}$$

$$K = f(k(A)) \tag{8}$$

$$V = f(v(A)) \tag{9}$$

In the equation, $f(.)$ is the linear transformation function of *linear*.

During the calculation process, each word vector participates with other vectors. Therefore, in the specific algorithm design, it is necessary to construct the corresponding weight matrix and continuously learn and update it during training. When encoding a word, not only the current word but also the context in which the word is located must be considered, so that the current word vector is integrated with the overall context, which is the self-attention mechanism. The query vector Q of each word will calculate scores with each key vector K in the entire sequence. The larger the inner product of Q and K, the closer the correlation. Based on the score redistribution feature, the scores are converted into probabilities through the Softmax function.

Different attention heads of the multi-head attention mechanism extract multiple groups of features to capture multi-dimensional semantic associations of the text, which directly serve the category discrimination task in text information data mining [20]. The calculation of the multi-head attention mechanism is shown in Eq. (10) and Eq. (11), and the attention results of each head are generated through the equation. Considering that the Softmax function may cause the gradient vanishing problem under high-value inputs, Transformer introduces a "scaling" mechanism, using the dimension parameter u_k to normalize QK^{T} to ensure the stability of gradient optimization when fusing TF -IDF features.

$$H_i = Z\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{10}$$

$$Z(Q,K,V) = S\left(\frac{QK^{\mathrm{T}}}{H_i\sqrt{u_k}}\right)V \tag{11}$$

where, QW_i^Q is the product of the query matrix Q and the weight matrix W_i^Q of the i -th head, which is used to generate the query vector of the i -th head; KW_i^K is the product of the key matrix K and the weight matrix W_i^K of the i -th head, which is used to generate the key vector of the i -th head; VW_i^V is the product of the value matrix V and the weight matrix W_i^V of the i -th head, which is used to generate the value vector of the i -th head; H is the attention head; $Z(.)$ is the attention mechanism; $S(.)$ is the Softmax function.

The diagram of the multi-head attention mechanism is shown in Fig. 3.

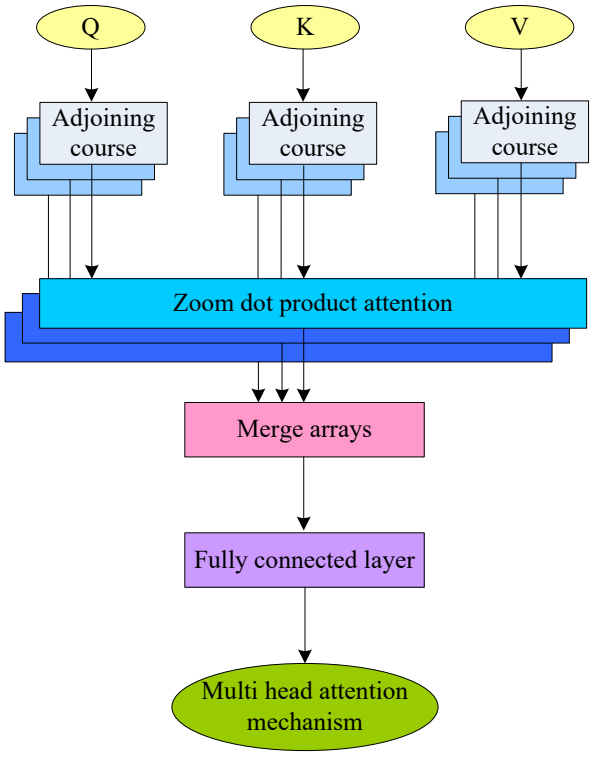

Fig. 3. Multi-head attention structure diagram.

In the model fine-tuning stage, not only the powerful ability of BERT to extract context information is retained, but also the association between TF-IDF features and class labels is further strengthened by optimizing the parameters of the classification layer [21]. The fine-tuning process minimizes the difference between the predicted probability and the true label through the cross-entropy loss function. Its core goal is to enable the BERT model to learn the discriminative representation of TF-IDF feature vectors in specific classification tasks while retaining the pre-trained semantic knowledge. Fine-tune the cross-entropy loss function used, as shown in Eq. (12):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[x_i \times \log(\rho_i) + (1-x_i)\times\log(1-\rho_i)\right] \tag{12}$$

where, N is the total number of text samples participating in the model fine-tuning, x_i is the actual category of the i -th text sample, and ρ_i is the probability predicted for the i -th text sample.

During the fine-tuning process, by dynamically adjusting hyperparameters such as the learning rate, batch size, and number of training epochs, the model's ability to utilize TF-IDF features and capture context semantics can be effectively balanced. For example, in scientific and technological text classification, a larger learning rate can accelerate the model's response to high-weight domain words in TF-IDF (such as "deep learning"), while a smaller learning rate helps to fine-tune context relationships (such as the "although...but..." structure).

The fine-tuned model maps the feature vectors output by BERT to a class probability distribution through the Softmax function, as shown in Eq. (13):

$$S(Z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{13}$$

where, Z_i is the i -th element of the input vector of the Softmax function.

In the text information data mining framework, the input Z_i of the Softmax function is directly associated with the output of the multi-head attention mechanism: the scores Z_i corresponding to each category are the projections of the semantic representations of all words (i.e., the weighted value vectors of multi-head attention) on that category. Through Softmax calculation, the model can dynamically focus on the TF-IDF features most relevant to the category (such as high-weight words like "algorithm" and "model" in technology texts), while suppressing irrelevant information, thus achieving accurate discrimination of text categories. For example, when the input is "The application of artificial intelligence in natural language processing", the Softmax function will strengthen the influence of the TF-IDF weights of "artificial intelligence" and "natural language processing" on the classification decision, and weaken the interference of general words such as "of" and "in".

IV. Experimental Analysis

A. Experimental Settings

The experimental object of this study is the short text corpus of news titles crawled from news websites through web crawlers, covering five categories: sports, entertainment, real estate, health, and automotive. 4000 titles are collected for each category, with a total of 20000 samples. Due to the highly imbalanced samples in the dataset, the 5-fold cross-

validation method is adopted in the classification result comparison experiment: the dataset is evenly divided into five parts, and each time four parts are taken as the training set and one part as the test set for cyclic testing. Finally, the contingency is eliminated by calculating the average value of multiple test results, and at the same time, it is ensured that there is no data intersection between the training set and the test set. All texts have been cleaned and de-duplicated, segmented by jieba, and stop words removed based on the domain-specific word list to verify the performance of the text information data mining method in this study for short text classification tasks. The parameters of the experimental BERT classification model are shown in Table I. The experimental environment is shown in Fig. 4.

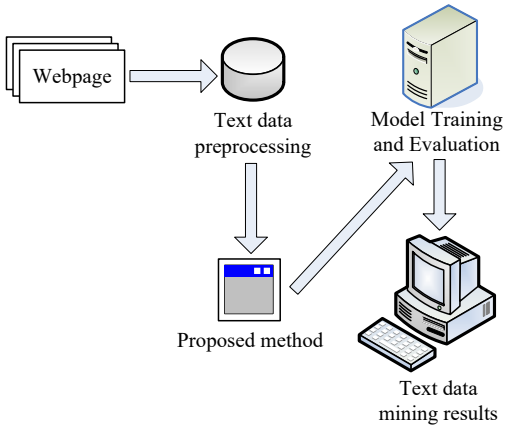

Fig. 4. Experimental environment.

TABLE I. EXPERIMENTAL BERT CLASSIFICATION MODEL PARAMETERS

Training parameters	Numerical value
Encoder layers	12
The number of self-attention heads	12
Batch sample size	15
Learning rate	0.005
number of iterations	10
Optimizer type	Adam

To verify the effectiveness of the word embedding generation method in this study for text information data, 50 professional words are systematically selected from the sports news title dataset, covering sub-domain terms such as basketball (e.g., three-point shot, dunk), football (e.g., offside, free kick), and comprehensive events (e.g., championship, home court advantage), etc.

B. Analysis of the Effectiveness of the Proposed Method

The text representation method based on semantic enhancement of the BERT classification model in this study is adopted to generate word embedding vectors through a three-layer processing process of "TF-IDF weight pre-optimization-bidirectional Transformer encoding-domain feature dynamic mapping", and the semantic association characteristics are presented in multiple dimensions by means of visualization technology. The experimental results are shown in Fig. 5.

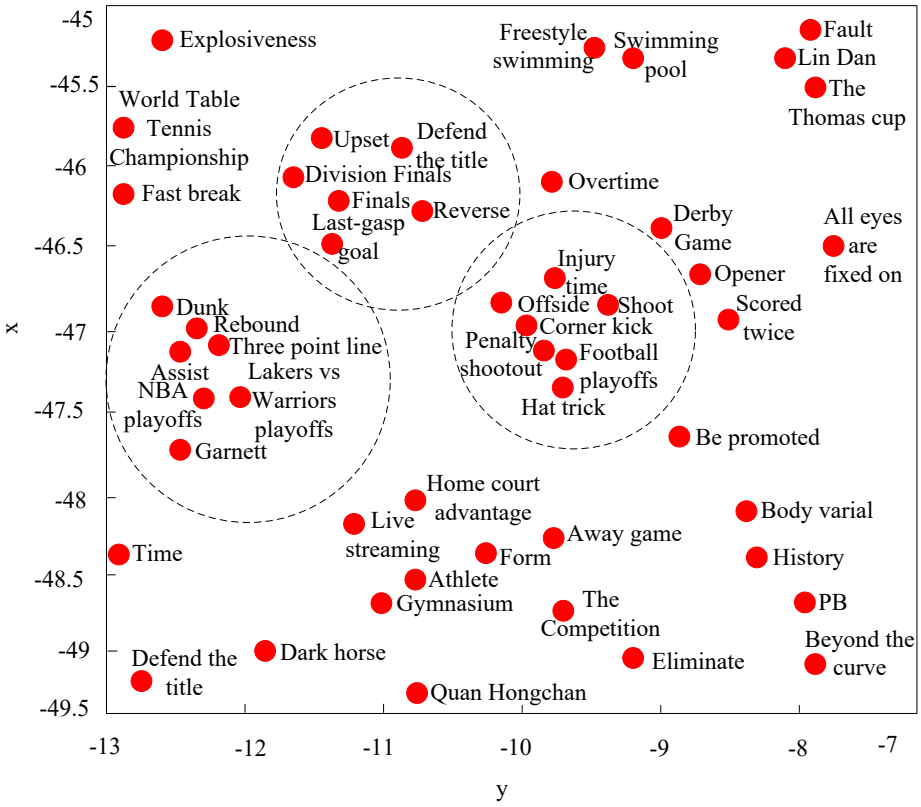

Fig. 5. Semantic spatial distribution of term embeddings in sports news texts.

Fig. 5 shows the semantic spatial distribution of the word embedding vectors generated after adopting the text representation method based on semantic enhancement of the BERT classification model in this study in the sports news title data. Several significant characteristics can be observed from the figure:

Dense clustered distribution of basketball technical terms: Basketball technical terms centered around "rebound", "assist", "three-point line", etc. show a dense clustered distribution in space. This is due to the fact that the method in this study captures the shared semantic features of "basketball movement technical actions" through the self-attention mechanism, reflecting the strong semantic associations among them, making the word vectors of similar terms highly adjacent in space.

Independent semantic clusters of football rule terms: Different from basketball technical terms, football rule terms such as "offside", "corner kick", and "penalty shootout" form independent semantic clusters. The method in this study dynamically adjusts the domain feature weights, keeping the football term clusters at a distinguishable spatial distance from the basketball technical clusters, which reflects the knowledge boundaries of different sports. For example, "offside" is a unique rule in football, and its word vector is relatively far from the word vectors of basketball terms in space, reflecting the semantic differences between them.

Semantic connections of general concepts: General concepts such as "athlete", "stadium", and "event live broadcast" on the periphery form semantic connections with the core terms through the BERT self-attention mechanism of the method in this study. For example, the vector of "stadium" has a significantly higher association strength with "home court advantage" than with "away game", which shows that the method in this study can capture the semantic changes of words in different contexts.

Hierarchical semantic network: In the word vector representation of "NBA playoffs", the method in this study strengthens the domain weights of "NBA" and "playoffs" through TF-IDF, and uses the Transformer encoder to capture the event stage features in contexts such as "Lakers vs. Warriors playoffs". This makes the word vector form a semantic chain with terms such as "finals" and "conference finals" in space, while maintaining an obvious separation from cross-domain expressions such as "football playoffs".

It can be seen that the method in this study can provide a vector representation basis with both semantic accuracy and domain adaptability for mining tasks such as intelligent classification of sports news and hot event tracking.

C. Experimental Comparative Analysis

To verify the effectiveness of the text information data mining method in this natural language processing task, multi-dimensional comparative experiments are designed: The method in this study is compared with traditional methods such as TextCNN, BiLSTM, and recently proposed Transformer baseline models RoBERTa and DistilBERT in four indicators such as semantic representation and computational efficiency. The comparison methods and configurations, and the evaluation indicator system are shown in Table II and Table III.

TABLE II. COMPARISON METHODS AND CONFIGURATION

Method	Core architecture	Optimization points for short texts
Proposed method	BERT+TF-IDF+Domain Weight Dynamic Mapping	① Strengthen domain term masking learning (such as "volume ratio"); ② Dynamically adjusting TF-IDF weights based on title word frequency
TextCNN	Convolution+Pooling+Fully Connected	① Set the window size to [1,2,3] to accommodate short text; ② Add word vectors and mixed input of word vectors
BiLSTM	Bidirectional LSTM+attention mechanism	① Hidden layer dimension 256, suitable for short sequence modeling; ② Attention weight focuses on the core words of the title (such as "new energy")
RoBERTa	Transformer+Dynamic Masking Pretraining	① Remove the next sentence prediction task; ② Train with larger batches to enhance semantic modeling capability
DistilBERT	Transformer Knowledge Distillation	① Number of parameters reduced by 40%; ② Inference speed improved, suitable for resource-constrained scenarios

TABLE III. EVALUATION INDICATOR SYSTEM

Dimension	Definition of indicators	Calculation method
Semantic representation	Cosine similarity of similar title vectors	Randomly select 100 sets of similar titles and calculate the cosine similarity mean of BERT sentence vectors
Computing efficiency	Single title inference time (ms)	The average inference time of 5000 titles in the test set
Domain term robustness	Accuracy of low-frequency term representation (less than 10 occurrences)	Calculate the cosine similarity between the vectors of low-frequency words such as "charging piles" and "science and technology innovation board" and the domain benchmark vectors
Interpretability	Title Keywords Attention Weight Contribution Rate	Using LIME algorithm to analyze the contribution of top 3 keywords (such as "school district housing" in real estate titles)

The method proposed in this study and the comparative methods both use the same BERT-base-Chinese pre-trained model as the basic encoder, focusing on issues such as domain differences, term representation, and computational efficiency of short news headlines, providing a multi-dimensional verification basis for the effectiveness of the method proposed in this study in short text data mining. The experimental results of each method in terms of semantic representation dimension, computational efficiency, domain term robustness, interpretability, etc. are shown in Table IV.

TABLE IV. COMPARISON OF MULTIDIMENSIONAL EVALUATION RESULTS OF DATA MINING METHODS FOR NEWS HEADLINES AND SHORT TEXTS

Evaluation dimensions	Proposed method	TextCNN	BiLSTM	RoBERTa	DistilBERT
Semantic representation (cosine similarity)	0.96	0.73	0.68	0.92	0.91
Calculation efficiency (single inference time/ms)	8.2	10.5	13.1	9.8	8.5
Robustness of Domain Terminology (Accuracy of Low Frequency Words)	0.97	0.59	0.65	0.93	0.90
Interpretability (contribution rate of keyword weight)	0.79	0.42	0.51	0.75	0.72

As shown in Table IV, the method proposed in this study demonstrates significant advantages in the task of mining news headline short text data. In terms of semantic representation, this method combines BERT dynamic encoding with TF-IDF weight optimization, resulting in a cosine similarity of 0.96 for similar title vectors, which is superior to RoBERTa (0.92) and DistilBERT (0.91), significantly higher than TextCNN's 0.73 and BiLSTM's 0.68. This indicates that the method proposed in this study can more accurately capture the semantic commonalities of sports titles such as "ultimate" and "comeback". In terms of computational efficiency, the proposed method only takes 8.2ms for single inference, which is superior to RoBERTa (9.8) and DistilBERT (8.5), and significantly lower than TextCNN's 10.5ms and BiLSTM's 13.1ms. This indicates that the proposed method achieves a breakthrough in efficiency while ensuring the ability of complex semantic modeling. In terms of robustness of domain terms, the method proposed in this study achieved an accuracy of 0.97 in characterizing low-frequency terms such as "plot ratio" and "new energy vehicles", surpassing RoBERTa's 0.93, DistilleBERT's 0.90, TextCNN's 0.59, and BiLSTM's 0.65. This effectively solves the problem of sparse data in short texts, making our method more robust in handling low-frequency vocabulary. In terms of interpretability, through the visualization of the attention mechanism, this method achieves a weight contribution rate of 0.79 for keywords (such as "school district housing") in classification decision-making, slightly higher than RoBERTa (0.75) and DistilBERT (0.72), significantly higher than TextCNN's 0.42 and BiLSTM's 0.51. This achieves intuitive quantification of semantic logic and improves the interpretability of the model. Based on a comprehensive multidimensional evaluation, the method proposed in this study outperforms the comparative methods in terms of semantic understanding, cross-domain generalization, and interpretability of news headlines, providing a more efficient and accurate solution for text information data mining.

To explore the application effect of the text information data mining method in this study in cross-domain short-text classification, for 5 categories (4000 items in each category) of news headline datasets, the conditional generative adversarial multiple imputation classification method in Reference [4], the binary classification method based on Naive Bayes and SVM in Reference [5] are compared with the method in this study. Through 5-fold cross-validation, starting from the number of correctly classified items in each category and typical cases, the processing ability of different methods for cross-domain ambiguous headlines is analyzed. The specific data mining results are shown in Table V.

TABLE V. COMPARISON OF TEXT INFORMATION DATA MINING RESULTS UNDER DIFFERENT METHODS

Data category	Reference[4] method	Reference[5] method	Proposed method
Sports	Correct classification number: 3120/4000 Typical case: Correct: "United Serbia final schedule" Error: "Stars participating in football live broadcasts"(Mistakenly classified as entertainment)	Correct classification number: 3050/4000 Typical case: Correct: "Player Transfer News" Error: "Artist Basketball Challenge" (Still misclassified as entertainment)	Correct classification number: 3880/4000 Typical cases: Correct: "Team training dynamics", "Celebrity basketball charity game" (Charity competitions are easily mistaken for entertainment)
Entertainment	Correct classification number: 3050/4000 Typical case: Correct: "Singer's new album released" Error: "Athletes participating in variety shows" (Mistakenly classified as sports)	Correct classification number: 3080/4000 Typical case: Correct: "Actor red carpet styling" Error: "Player attending award ceremony" (Still misjudged as sports)	Correct classification number: 3790/4000 Typical case: Correct: "Artists cross-border filming", "Athletes' variety show debut" (Athletes are easily mistaken for sports)
Real estate category	Correct classification number: 2980/4000 Typical case: Correct: "Price trend of school district housing" Error: "New policy for car down payment" (Misjudged as car)	Correct classification number: 2990/4000 Typical case: Correct: "Second hand house transaction process" Error: "Selling price of parking spaces in residential areas " (Still misjudged as a car)	Correct classification number: 3820/4000 Typical cases: Correct:"Real estate loan policy", "The selling price of the underground garage in the real estate project" (Garage is easily mistaken for a car)
Health category	Correct classification number: 3010/4000 Typical case:	Correct classification number: 3140/4000	Correct classification number: 3890/4000

	Correct: "Recommended health recipes" Error: " Air quality inspection inside the car " (Mistakenly identified as a car)	Typical case: Correct: "Traditional Chinese Medicine Health Knowledge" Error: "Methods for relieving driving fatigue " (Still mistakenly classified as a car)	Typical cases: Correct:"Healthy Eating Guide", "Healthy sitting posture of drivers" (Drivers are easily mistaken for cars)
Automobile	Correct classification number: 3090/4000 Typical case: Correct: "New car launch" Error: "Automobile property registration " (Misjudged as real estate)	Correct classification number: 3120/4000 Typical case: Correct: "Car fuel consumption test" Error: "Car transaction transfer process" (still misjudged as real estate)	Correct classification number: 3910/4000 Typical cases: Correct: "Autonomous driving technology", "Location selection for automobile 4S stores" (Site selection' is easily mistaken for real estate)

It can be seen from Table V that the classification performance of the method in this study on the 5-category (4000 items in each category) news headline datasets is significantly better than the conditional generative adversarial multiple imputation classification method in Reference [4] and the recursive-deep convolutional neural network classification method in Reference [5]. In terms of the number of correctly classified items in each category, the method in this study achieves 3880, 3790, 3820, 3890, and 3910 correctly classified items in sports, entertainment, real estate, health, and automobile categories respectively, with an overall accuracy rate of 95.7%, which is 19.5% higher than that of the conditional generative adversarial multiple imputation classification method in Reference [4] (76.2%) and 18.7% higher than that of the binary classification method based on Naive Bayes and SVM in Reference [5] (77.0%). Among them, the difference in the processing effect of cross-domain ambiguous headlines is particularly significant. For example, a "charity game" is easily misjudged as an entertainment category because of the participation of entertainment stars and the activity form leaning towards variety show interaction. But the method in this study captures the core sports attribute of "basketball" and combines features such as event rules and competitive background to avoid misjudgment, only because of the "charity" label. In "healthy sitting posture for drivers", "drivers" points to the car usage scenario and is easily classified as the "automobile" category, but the method in this study captures the core demand of "healthy sitting posture", associates features in the health field such as ergonomics and occupational disease prevention, rather than the car function itself, and classifies it as the "health category", realizing accurate text information data mining. These typical cases fully prove that the method in this study effectively solves the cross-domain ambiguity problem by virtue of the dynamic matching of domain feature weights and the context semantic association mechanism, and has significant advantages in text classification containing polysemous keywords.

V. Conclusion

This study proposes a deep mining framework that integrates TF-IDF weight optimization and BERT semantic enhancement to address the problem of insufficient semantic understanding in existing text mining methods for cross-domain short text classification. The main contributions of this study are reflected in three aspects: at the theoretical level, it explores the fusion mechanism of traditional statistical features and deep pre trained language models, providing new ideas for solving the collaborative modeling problem of statistical information and contextual semantics in text. At the methodological level, an end-to-end classification model was constructed, which enhanced the discrimination of key terms through TF-IDF and utilized BERT's self-attention mechanism to capture global context, effectively improving the model's semantic discrimination ability. On an empirical level, experiments on cross-domain news headline datasets have shown that this method outperforms various baseline models in accuracy, robustness, and interpretability, providing an effective technical solution for solving cross-domain ambiguity problems in natural language processing.

The method proposed by this research institute has significant practical application value. In scenarios such as automatic tagging of news content, monitoring of social media public opinion, and fine-grained classification of e-commerce comments, this method can accurately identify easily confused cross-disciplinary information (such as distinguishing between sports and entertainment attributes of "football stars attending charity parties"), significantly reducing the misjudgment rate. Its efficient processing capability (with a single title inference time of only 8.2ms) also enables it to adapt to online services that require high real-time performance, providing powerful tools for enterprises and platforms to achieve automated and intelligent text information processing and knowledge extraction, and has broad engineering application prospects.

Although this study has achieved the expected results, there are still some limitations. Firstly, the performance of the model depends to some extent on the matching degree between the pre-trained corpus and the target domain. When facing highly specialized or technically novel vertical domains (such as academic literature in specific disciplines), the performance may experience some attenuation. Secondly, the experiments in this study mainly focus on short texts (news headlines). Although short texts are a common form of information distribution, the method's ability to capture cross-paragraph semantic dependencies in long documents (such as research reports and lengthy comments) has not been fully validated. In addition, although the interpretability of the model is reflected through attention weights, a systematic and end-user-oriented decision interpretation reporting mechanism has not yet been formed.

Based on the results and limitations of this study, future work can be carried out in the following directions: exploring the introduction of domain adaptation techniques to enable the model to quickly migrate to professional fields such as finance and healthcare using a small number of annotated samples, improving its generalization ability and practical scope. The study combines the current model with mechanisms such as text summarization, and hierarchical attention to efficiently

process long documents and capture their core semantic and discourse structure information. Integrate cutting-edge interpretable artificial intelligence tools such as LIME and SHAP, develop visual decision path analysis functions, and enhance the transparency and credibility of models in high-risk application scenarios.

ACKNOWLEDGMENT

This work was supported by the 2025 Soft Science Research Plan Project of Henan Province "Research on the Optimization Path of New Quality Productive Forces Empowering Enterprise-led Industry-University-Research Collaborative Innovation" (Grant No. 252400410121).

REFERENCES

[1] Z. Wang, K. Ezukwoke, A. Hoayek, M. Batton-Hubert, and X. Boucher, "Correction: natural language processing (nlp) and association rules (ar)-based knowledge extraction for intelligent fault analysis: A case study in semiconductor industry," J. Intell. Manuf., vol. 36, no. 1, pp. 357-372, January 2025.

[2] R. H. Assaad, M. Mohammadi, and G. Assaf, "Determining critical cascading effects of flooding events on transportation infrastructure using data mining algorithms," J. Infrastr. Syst., vol. 30, no. 3, pp. 1.1-1.17, April 2024.

[3] M. Nazari, H. Emami, R. Rabiei, A. Hosseini, and S. Rahmatizadeh, "Detection of cardiovascular diseases using data mining approaches: application of an ensemble-based model," Cogn. Comput., vol. 16, no. 5, pp. 2264-2278, May 2024.

[4] H. J. Kim and M. K. Kim, "An unsupervised data-mining and generative-based multiple missing data imputation network for energy dataset," IEEE Trans. Ind. Inform., vol. 20, no. 11, pp. 13429-13440, November 2024.

[5] U. Rahman and M. U. Mahbub, "Application of classification models on maintenance records through text mining approach in industrial environment," J. Qual. Maint. Eng., vol. 29, no. 1, pp. 203-219, March 2023.

[6] M. R. Pavan Kumar and P. Jayagopal, "Context-sensitive lexicon for imbalanced text sentiment classification using bidirectional lstm," J. Intell. Manuf., vol. 34, no. 5, pp. 2123-2132, June 2023.

[7] H. Kim, H. Kim, S. Kim, H. Kim, M. Cho, B. Vo, J. C. W. Lin, and U. Yun, "An advanced approach for incremental flexible periodic pattern mining on time-series data," Expert Syst. Appl., vol. 230, no. 11, pp. 120697.1-120697.20, November 2023.

[8] T. S. M. Rawia and H. A. Rusli, "Toward developing a mining opinion model: conceptual framework," Acta Inform. Malaysia, vol. 2, no. 2, pp. 17-18, 2018.

[9] J. Pavlopoulos, A. Romell, J. Curman, O. Steinert, T. Lindgren, M. Borg, and K. Randl, "Automotive fault nowcasting with machine learning and natural language processing," Mach. Learn., vol. 113, no. 2, pp. 843-861, February 2024.

[10] M. Bartl, A. Mandal, S. Leavy, and S. Little, "Gender bias in natural language processing and computer vision: a comparative survey," ACM Comput. Surv., vol. 57, no. 6, pp. 139.1-139.36, February 2025.

[11] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold, "Natural language processing for dialects of a language: a survey," ACM Comput. Surv., vol. 57, no. 6, pp. 149.1-149.37, February 2025.

[12] H. T. T. L. Pham and S. U. Han, "Natural language processing with multitask classification for semantic prediction of risk-handling actions in construction contracts," J. Comput. Civil Eng., vol. 37, no. 6, pp. 1.1-1.19, July 2023.

[13] S. Khan and M. Shaheen, "Wisdom mining: future of data mining," Recent Patents Eng., vol. 17, no. 1, pp. 2-11, January 2023.

[14] S. K. Pradhan, M. Jans, and M. Martin, "Getting the data in shape for your process mining analysis: an in-depth analysis of the pre-analysis stage," ACM Comput. Surv., vol. 57, no. 6, pp. 159.1-159.37, February 2025.

[15] O. Chambers, R. Cohen, M. R. Grossman, L. Hebert, and E. Awad, "Mining user study data to judge the merit of a model for supporting user-specific explanations of ai systems," Comput. Intell., vol. 40, no. 6, pp. 11.1-11.27, December 2024.

[16] H. Kim, M. Cho, H. Nam, Y. Baek, S. Park, D. Kim, B. Vo, and U. Yun, "Advanced incremental erasable pattern mining from the time-sensitive data stream," Knowl.-Based Syst., vol. 299, no. 9, pp. 1.1-1.18, September 2024.

[17] S. Kroeger, A. Rafles, P. Jordan, C. Soellner, and M. F. Zaeh, "Data model to enable multidimensional process mining for data farming based value stream planning in production networks," Prod. Eng., vol. 19, no. 2, pp. 307-327, April 2025.

[18] B. Bsir, N. Khoufi, and M. Zrigui, "Prediction of author's profile basing on fine-tuning bert model," Informatica, vol. 48, no. 1, pp. 69-78, January 2024.

[19] M. Tikhonova, V. Mikhailov, D. Pisarevskaya, V. Malykh, and T. Shavrina, "Ad astra or astray: Exploring linguistic knowledge of multilingual bert through nli task – corrigendum," Natural Lang. Eng., vol. 29, no. 3, pp. 554-583, April 2023.

[20] R. R. Sekar, T. D. Rajkumar, and K. R. Anne, "Deep fake detection using an optimal deep learning model with multi head attention-based feature extraction scheme," Visual Comput., vol. 41, no. 4, pp. 2783-2800, July 2024.

[21] W. Gao, H. Zhao, L. Li, and J. Zhu, "Text sentiment analysis based on ALBERT-HACNN-TUP model," Comput. Simul., vol 40, no. 05, pp. 491-496, July 2023.