Enhanced LLM Text Classification Method with Embedded Semantic Feature Encoding

Meng Wang[1], Jing Xie[1], Yang Li[1,2], Zhixiong Zhang[1,2] and Hanyu Li[1,2]
[1]*National Science Library, Chinese Academy of Science, Beijing, China*
[2]*University of Chinese Academy of Science, Beijing, China*

Keywords: Text Classification, Semantic Feature Encoding, Large Language Models, Feature Embedding.

Abstract: Accurate identification of semantic features in scientific texts is crucial for enhancing text classification performance. This paper presents a large language model text classification method with embedded semantic feature encoding, which enhances the model's understanding of textual semantics through a dual semantic feature encoding mechanism. The method employs a dynamic window-based local-global feature extraction strategy to capture topical semantic features and utilizes hierarchical structural aggregation mechanisms to extract organizational semantic information from texts. To fully leverage the extracted semantic features, we design a feature replacement encoding strategy that embeds topical semantic features and structural semantic features into the [CLS] and [SEP] positions of large language models, respectively, achieving deep fusion between semantic features and internal model representations, thereby improving the accuracy and robustness of text classification. Experimental results demonstrate that the proposed semantic feature encoding enhancement method achieves significant performance improvements. On the DBPedia dataset, the semantically encoded SciBERT model achieves an F1-score of 91.07%, representing a 5.26% improvement over the original encoding approach. In the scientific literature value sentence identification task, Qwen3-14B combined with semantic feature encoding and QLora fine-tuning achieves an F1-score of 94.19%, showing a 14.64% improvement over the baseline model. Compared to traditional feature concatenation or simple fusion approaches, our feature replacement encoding strategy leverages semantic features at critical positions, significantly enhancing both classification precision and recall. Ablation experiments further validate the synergistic effects of topical semantic features and structural semantic features, confirming the effectiveness of the dual semantic feature encoding mechanism. The research findings highlight the advantages of semantic feature encoding in text classification tasks, providing an effective technical solution for intelligent analysis of scientific texts.

1 INTRODUCTION

Text classification is a fundamental task in natural language processing that aims to accurately categorize text into predefined multiple classes based on semantic content and expressive features (Feng et al., 2023). Text multi-classification plays a crucial role in semantic structure parsing, key information extraction, and knowledge organization of scientific and technical texts (Li et al., 2024). Therefore, optimization and improvement of text multi-classification models have consistently been one of the research hotspots in the field of natural language processing (Wang et al., 2025). Semantic features, as key representations that characterize the core content and deep meaning of texts, play an important role in improving the accuracy and efficiency of text multi-classification models (Liu et al., 2024). However, with the expansion of scientific text data scale and the increasing complexity of category systems, semantic boundaries between different text categories have gradually become blurred, and category distributions often exhibit significant imbalanced characteristics. This leads to difficulties for models in accurately capturing semantic differences between categories, particularly resulting in lower recognition accuracy for categories with fewer samples. Therefore, embedding semantic features of scientific texts into multi-classification models is of great significance.

Multi-classification models that consider text semantic features are primarily studied based on pre-trained models (PLMs) with feature word

Wang, M., Xie, J., Li, Y., Zhang, Z. and Li, H.
Enhanced LLM Text Classification Method with Embedded Semantic Feature Encoding.
DOI: 10.5220/0013697900004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 87-97

embeddings. PLMs obtain distributed representations of feature words in texts through training on large-scale corpora, and utilize attention mechanisms to calculate the correlation between feature words and their contexts to capture semantic features of texts (Devlin et al., 2019; Ezugwu et al., 2024). This feature word-based representation learning approach can effectively improve model performance in multi-classification tasks, but often treats feature words as independent semantic units, neglecting the semantic associations between feature words (Zhao et al., 2015; Pangakis & Wolken, 2024). To enhance the model's capability for text semantic understanding, researchers have attempted to integrate semantic information of texts through feature word semantic fusion mechanisms, such as designing feature word attention modules or constructing graph structure-based semantic dependency relationships (Kokkodis et al., 2025; Wang et al., 2024). However, existing feature word semantic fusion strategies mostly adopt unified processing approaches, failing to fully consider the differentiated contributions of various text features in classification tasks, which may lead to models' inability to accurately identify subtle semantic differences between categories.

With the rapid development of Large Language Models (LLMs) in the field of natural language processing, their strong semantic understanding and knowledge representation capabilities provide new insights for addressing the problem of semantic feature utilization in text multi-classification (Achiam, 2023; Touvron et al., 2023). Through pre-training on massive text data, LLMs can not only capture topical semantic features of texts but also deeply understand structural semantic information of texts, laying a foundation for achieving more comprehensive semantic representation. Meanwhile, LLMs possess fine-grained semantic discrimination capabilities that can identify unique semantic features of texts from different categories, which provides possibilities for solving class distribution imbalance problems (Guo et al., 2024). However, how to effectively integrate the semantic understanding advantages of LLMs and achieve accurate recognition of imbalanced categories based on this foundation remains a key challenge facing text multi-classification tasks.

To address these challenges, this paper proposes a LLMs text classification method with embedded semantic feature encoding. This method leverages the semantic understanding capabilities of LLMs to construct a dual semantic encoding mechanism, directly embedding topical semantic and structural semantic features into the model's internal architecture. Through a feature replacement encoding strategy, the method enhances the model's understanding of holistic textual semantics, achieving deep fusion between semantic features and contextual content, thereby significantly improving the accuracy and robustness of text classification.

The main contributions of this paper are summarized as follows:

- To enhance the semantic understanding capability of LLMs in text classification tasks, a dual semantic feature encoding mechanism is proposed, which extracts topical semantic features through dynamic window-based local-global feature extraction and captures structural semantic features via hierarchical aggregation mechanisms.
- To achieve deep integration between semantic features and model internal representations, a feature replacement encoding strategy is designed that directly embeds topical semantic features and structural semantic features into the [CLS] and [SEP] positions of LLMs, respectively.
- To address the challenge of semantic boundary ambiguity in scientific text classification, particularly the performance degradation on imbalanced datasets, the proposed method leverages the semantic discrimination capabilities of LLMs combined with explicit semantic feature encoding.
- To validate the effectiveness and generalizability of the proposed semantic feature encoding method, comprehensive experiments are conducted on both public benchmark datasets and domain-specific scientific literature datasets, covering multiple PLMs and LLMs.

2 LITERATURE REVIEW

2.1 Text Classification By PLMs

In text classification research, scholars primarily conduct studies from two perspectives: PLMs-based text classification and LLMs-based text classification. In PLMs-based text classification research, the main approach involves fine-tuning PLMs such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019), fully leveraging their semantic representation capabilities learned on large-scale

corpora to improve classification performance. Li et al. (2022) proposed a research question sentence recognition model based on BERT-CNN (Convolutional Neural Networks), using BERT as a word embedding model and then utilizing CNN to extract linguistic information features at different levels from sentence vectors to achieve sentence classification. Experimental results demonstrated that BERT's text representation capability contributes to improving accuracy in sentence classification tasks. In 2019, Liu et al. (2019) proposed an improved BERT model, namely RoBERTa, which learns language nuances and complexities based on larger training datasets and longer training time during pre-training and removes the NSP (Next Sentence Prediction) task. Liu and Cao (2022) respectively used RoBERTa, BERT, and BiLSTM-CRF (Bidirectional Long Short-Term Memory-Conditional Random Field) to identify named entities in winter sports text news. Results showed that compared to BERT and BiLSTM-CRF, RoBERTa's F1 score improved by 0.77% and 1.81%, respectively. Liu et al. (2022) proposed a causal intervention-based method for weakening confounding factors based on the RoBERTa model, thereby improving the accuracy of few-shot relation classification tasks and reducing the impact of confounding factors on model performance. However, PLMs primarily rely on masked prediction and sentence-level pre-training tasks, showing limitations in capturing fine-grained semantic features and handling complex contextual relationships.

2.2 Text Classification by LLMs

Compared to PLMs, LLMs possess stronger semantic understanding and knowledge transfer capabilities, providing new research directions for optimizing text classification tasks. LLMs-based text classification research mainly includes two categories: (1) Instruction fine-tuning-based classification methods enhance models' understanding of classification tasks by designing specific prompt templates. Han (2023) proposed fine-tuning based on the Qwen-7B model for financial text classification problems. Experimental results showed that the fine-tuned LLMs achieved an accuracy of 82.27%, surpassing Deberta-V3-base and Deberta-V3 models, indicating that fine-tuned LLMs can enhance text classification effectiveness. Zhang et al. (2024) proposed an adaptive text classification enhancement framework that combines a specialized text classification model by adjusting training sample distributions and iteratively fine-tuning LLMs. Chae and Davidson (2024) explored the application of LLMs in supervised text classification, comparing prompt-based zero-shot and few-shot learning, fine-tuning with more training data, and instruction fine-tuning that combines prompts and training data. They found that instruction fine-tuning approaches are beneficial for improving models' text classification capabilities for complex text classification tasks. Fatemi et al. (2025) employed model merging techniques, integrating single-task domain-specific fine-tuned models with base models for financial domain text classification.

However, instruction fine-tuning-based LLMs text classification tends to learn features of mainstream categories with larger sample sizes, with limited recognition capability for minority categories. Particularly, when different categories exhibit semantic differences in expression, simple prompt templates struggle to accurately capture subtle distinctions between categories, leading to constrained classification performance on imbalanced datasets. Therefore, to address the data imbalance problem, (2) data augmentation-based classification methods focus on enhancing models' learning capabilities for category features. Peng and Shao (2024) proposed a data augmentation framework targeting sample imbalance problems in text classification, adjusting instruction-to-data mappings obtained from classification datasets and generating data based on GPT-4. Experimental results demonstrated that instruction fine-tuning based on generated data benefits model classification accuracy. Meguellati et al. (2025) utilized LLMs to clean noisy text and provide context-rich explanations, thereby enhancing training sets without substantially increasing data volume. Experimental results showed that zero-shot enhanced LLMs performed poorly in text classification tasks compared to supervised models. However, after integrating LLM-based semantic enhancement, their performance was comparable to methods relying on human-annotated data. Guo et al. (2024) proposed a method using LLMs as data annotators to augment few-shot data and utilized the augmented data to fine-tune PLMs and LLMs respectively. Experimental results demonstrated that RoBERTa models trained on GPT-4 augmented data exhibited superior or comparable performance compared to models trained solely on human-annotated data.

3 OVERALL FRAMEWORK

The overall framework of the LLM text classification method with embedded semantic feature encoding

comprising two core components: the semantic feature extraction module and the semantic feature encoding embedding module. In the semantic feature extraction module, we construct a dual semantic feature extraction mechanism that employs dynamic window-based local-global feature extraction methods to capture topical semantic features of texts, while simultaneously utilizing hierarchical structural feature extraction mechanisms to capture organizational semantic features of texts. In the semantic feature encoding embedding module, the extracted topical semantic features and structural semantic features are embedded into the [CLS] and [SEP] positions of LLMs through a feature replacement encoding strategy, achieving deep fusion between semantic features and contextual content, thereby enhancing the semantic representation of input texts. Finally, the semantically enhanced LLMs completes the text classification task and outputs classification results.

4 METHODOLOGY

4.1 Topical Semantic Feature Construction

Topical semantic features, as core elements of text classification, can effectively capture topical information and semantic content of texts, playing an important role in improving classification performance. Traditional topical feature extraction methods such as TF-IDF and LDA often require global computation and iterative optimization over entire documents, resulting in high computational complexity and difficulty in capturing fine-grained semantic associations within texts. Moreover, these methods only focus on global statistical information while neglecting semantic dependencies in local contexts, leading to limited feature expression capabilities. Therefore, this paper proposes a dynamic window-based local-global topical feature extraction method that improves feature expression capabilities in models while ensuring efficiency in semantic feature construction through the fusion of local and global features. Given input text sequence $X =< X_1, X_2, X_3, X_4 >$, the topical semantic feature construction process is as follows:

(1) Local topical feature extraction

To efficiently capture fine-grained semantic associations, a sliding window w is employed to scan the text and compute the topical aggregated representation of terms within the window:

$$h_{local}(i) = \sigma\left(\sum_{j=i-w/2}^{i+w/2} \alpha_j \cdot e_j\right) \quad (1)$$

where σ(·) is the activation function, w is the sliding window size, e_j represents the vector representation of the jth term, and α_j is the corresponding attention weight. The attention weights are calculated as following:

$$\alpha_j = softmax\left(e_j \cdot W_q \cdot e_i\right) \quad (2)$$

where W_q is the learnable query matrix parameter, e_i is the vector representation of the center term, and softmax is used to normalize the attention scores.

(2) Global topical relevance computation

To compensate for the limitation of considering only local information, a global topical representation of the text is constructed and interactively enhanced with local features:

$$h_{global} = Pool(h_{local}(i)_{i=1}^{n}) \quad (3)$$

where Pool(·) represents the pooling operation used to aggregate all local features to obtain a global representation. Therefore, the semantic feature$v_{topic}(i)$ extracted in this paper is represented as:

$$v_{topic}(i) = tanh\left(W_t \cdot \left[h_{local}(i); h_{global}\right] + b_t\right) \quad (4)$$

where W_t and b_t are the transformation matrix and bias term respectively, [;] denotes the feature concatenation operation, and tanh is the hyperbolic tangent activation function. Table 1 shows the pseudocode for topical semantic feature construction.

Table 1: The pseudocode for topical semantic feature construction.

Algorithm 1: Topic Semantic Feature Extraction with Dynamic Window.
Input: Text sequence X = {x_1, x_2, ..., x_n}, window size w Output: Topic semantic features v_topic 1: Initialize embedding matrix E = {e_1, e_2, ..., e_n} 2: Initialize query matrix W_q, transformation matrix W_t, bias b_t 3: for i = 1 to n do 4: start ← max(1, i - w/2) 5: end ← min(n, i + w/2) 6: for j = start to end do 7: α_j ← softmax(e_j · W_q · e_i) 8: end for 9: h_local(i) ← σ($\sum_{j=start}^{end}$ α_j · e_j) 10: end for 11: h_global ← Pool({h_local(i)}$_{i=1}^{n}$) 12: for i = 1 to n do 13: v_topic(i) ← tanh(W_t · [h_local(i); h_global] + b_t) 14: end for 15: return mean(v_topic)

4.2 Structural Semantic Feature Construction

Considering that different types of sentences have different writing structures, structural semantic features are crucial for understanding logical relationships between sentences in scientific texts. Traditional methods mainly rely on bag-of-words models or simple sequence models, which are inadequate for accurately capturing hierarchical structural relationships between sentences. To address this, this paper designs a lightweight structural feature extraction method that obtains structural semantic information of texts through hierarchical feature aggregation.

(1) Basic feature representation: For input text, first obtain the basic feature representation of each term:

$$h_{init}(i) = \text{FFN}([e_i; pos_i]) \quad (5)$$

where e_i is the vector representation of the term, pos_i is the positional encoding, *FFN* is the feed-forward neural network.

(2) Hierarchical structure aggregation: Based on dependency relationships between terms, perform layered feature aggregation:

$$h_l(i) = \sigma\left(\sum_{j \in N(i)} w_j \cdot h_{l-1}(j)\right) \quad (6)$$

where σ(·) is the activation function, $N(i)$ represents the set of adjacent terms to term i, w_j is the aggregation weight, l represents the layer number, and $h_{l-1}(j)$is the feature representation of the (l - 1)th layer.

(3) Inter-sentence relationship representation:

Construct sentence-level structural representation:

$$v_{struct} = \text{Pool}(\{h_L(i)\}_{i=1}^{n}) \quad (7)$$

where Pool(·) is the pooling operation, $h_L(i)$ is the term representation of the last layer, and v_{struct} is the final structural feature representation. Table 2 shows the pseudocode for structural semantic feature construction.

4.3 Semantic Feature Encoding Embedding Mechanism

To effectively embed the extracted semantic features into LLMs, this paper designs a feature replacement encoding mechanism. Traditional methods directly use PLMs' [CLS] and [SEP] tokens for text representation, where [CLS] is used to obtain global semantic representation and [SEP] is used to segment and mark sequence boundaries. Considering that topical semantic features v_{topic} also capture global semantic information of texts, while structural semantic features v_{struct} also contain boundary and organizational structure information of texts, these two types of features can functionally correspond to replacing [CLS] and [SEP] tokens respectively.The specific process of the semantic feature encoding embedding mechanism proposed in this paper is as follows:

Table 2: The pseudocode for structural semantic feature construction.

Algorithm 2: Hierarchical Structural Semantic Feature Extraction.
Input: Text sequence X = {x_1, x_2, ..., x_n}, number of layers L
Output: Structural semantic features v_struct
Initialize embedding matrix E = {e_1, e_2, ..., e_n}
Initialize position encoding pos = {pos_1, pos_2, ..., pos_n}
Initialize FFN, layer weights $\{W\_l\}_{l=1}^{L}$
for i = 1 to n do
h_init(i) ← FFN([e_i; pos_i])
end for
$h^{(0)} \leftarrow$ h_init
for l = 1 to L do
for i = 1 to n do
N(i) ← GetNeighbors(i)
$h^{(l)}(i) \leftarrow \sigma(\sum_{j \in n(i)} w\_j \cdot h^{(l-1)}(j))$
end for
end for
$v\_struct \leftarrow \text{Pool}(\{h^{(L)}(i)\}_{i=1}^{n})$
return v_struct

First, the original text sequence is passed through topical feature extraction and structural feature extraction modules to obtain corresponding semantic feature representations. Then, the extracted topical semantic features v_{topic} replace the [CLS] feature representation at the top of the sequence (purple block in the figure), while the structural semantic features v_{struct} replace the [SEP] feature representation at the bottom of the sequence (pink block in the figure), with the original text features in the middle remaining unchanged. The final fused feature sequence X_embed is shown as follows:

$$X_{embed} = [v_{topic}, X_1, X_2, X_3, X_4, v_{struct}] \quad (8)$$

Table 3 shows the pseudocode for semantic feature encoding embedding.

Table 3: The pseudocode for semantic feature encoding embedding.

```
Algorithm 3: Semantic Feature Encoding.
Input: Input sequence input_ids, attention mask
Output: Enhanced text representation
E ← WordEmbedding(input_ids)
v_topic ← TopicSemanticExtract(E)
v_struct ← StructuralSemanticExtract(E)
E_modified ← E.clone()
E_modified[0] ← v_topic      // Replace [CLS] token
sep_pos ← FindSEPPosition(input_ids)
E_modified[sep_pos] ← v_struct  // Replace [SEP] token
outputs ← Model(inputs_embeds=E_modified, attention_mask)
return outputs.pooler_output
```

5 APPROACH DESIGN

5.1 Topical Semantic Feature Construction

To validate the effectiveness of our proposed method, we designed a systematic experimental protocol comprising three phases:

(1) To verify the effectiveness of the feature encoding mechanism, we conducted experiments on the public DBPedia dataset (Lehmann et al., 2015) comparing traditional [CLS] and [SEP] encoding approaches with our proposed embedding of topical semantic features and structural semantic features, thereby validating the advantages of the feature fusion strategy.

(2) We performed ablation experiments on the same dataset to separately validate the contributions of topical semantic features and structural semantic features, demonstrating the necessity and complementarity of both feature types.

(3) To validate the effectiveness of LLMs binary classification with embedded universal semantic features, we conducted comparative performance experiments on our constructed scientific literature value sentence dataset to evaluate the performance advantages of our method.

Regarding model selection, to comprehensively validate the universality and effectiveness of our method, we selected BERT, SciBERT (Beltagy et al., 2019), and RoBERTa as baseline PLM models, and Qwen3-14B (Yang et al., 2025), LLaMa4-17B (Meta AI, 2025), and GLM4-9B (GLM et al., 2024) as baseline LLM models.

5.2 Datasets

To comprehensively evaluate the performance of the two-stage classification method enhanced with embedded semantic feature encoding for scientific texts, this paper selects the public dataset DBPedia and the constructed scientific literature value sentence dataset as experimental datasets to verify the model's effectiveness at different stages. The specific dataset information is as follows:

(1) DBPedia dataset. This dataset is a standard evaluation dataset for text classification tasks, derived from Wikipedia article abstracts and containing texts from 14 different thematic categories. The categories in the dataset cover 14 entity types including Company, Educational Institution, Artist, Athlete, etc. This paper selects DBPedia as a benchmark dataset for evaluating text classification model performance. With standardized text structure and clear themes, it can effectively verify the performance of the proposed feature encoding embedding mechanism and semantic feature extraction methods on public datasets.

(2) Scientific literature value sentence dataset. This dataset is used to evaluate the performance of scientific text value sentence recognition tasks, containing 23,912 scientific literature sentences composed of value sentences and non-value sentences with a positive-to-negative sample ratio of 1:1. The sentences in the dataset come from academic papers in fields such as computer science and engineering technology, annotated by professional annotators to form a high-quality binary classification dataset. This dataset serves as a professional dataset for evaluating value sentence recognition capabilities and can be used to verify the practical application effectiveness of the proposed method in the scientific text domain.

6 EXPERIMENTS

6.1 Semantic Feature Encoding Effectiveness Analysis

To verify the advantages of the proposed semantic feature encoding mechanism compared to traditional methods, we designed systematic comparison experiments on the DBPedia dataset. The experiment selected 20,000 texts each from Company and Educational Institution categories, forming a binary classification dataset with a 1:1 positive-to-negative sample ratio, totaling 40,000 samples. The dataset

Table 4: The results of semantic feature encoding effectiveness analysis.

Models	Encoding Method	A (%)	P (%)	R (%)	F1 (%)
BERT	Original Encoding	82.40	89.10	79.20	83.80
BERT	Semantic Encoding	**86.85 (+4.45)**	**91.20 (+2.10)**	**85.30 (+6.10)**	**88.17 (+4.37)**
SciBERT	Original Encoding	84.20	90.50	81.60	85.81
SciBERT	Semantic Encoding	**89.75 (+5.55)**	**92.80 (+2.30)**	**89.40 (+7.80)**	**91.07 (+5.26)**
RoBERTa	Original Encoding	83.60	89.80	80.40	84.85

was divided into training set (32,000 samples), validation set (4,000 samples), and test set (4,000 samples) in an 8:1:1 ratio. The experiment employed three pre-trained models—BERT, SciBERT, and RoBERTa—as basic architectures, comparing performance under both original encoding methods and the proposed semantic feature encoding methods. The original encoding method maintains the [CLS] and [SEP] tokens of pre-trained models unchanged, using traditional input sequence encoding; the semantic feature encoding method extracts topical and structural semantic features respectively and replaces the [CLS] and [SEP] position embedding vectors. All experiments used the same hyperparameter settings: learning rate of 1e-5, batch size of 32, 5 training epochs, and AdamW optimizer (Loshchilov & Hutter, 2019). The results of semantic feature encoding effectiveness analysis are shown in Table 4.

The results of semantic feature encoding effectiveness analysis are shown in Table 1. Experimental results demonstrate that the proposed semantic feature encoding method achieved significant performance improvements across all pre-trained models. On the BERT model, the semantic encoding method improved accuracy, precision, recall, and F1 score by 4.45%, 2.10%, 6.10%, and 4.37% respectively compared to the original encoding method. The SciBERT model showed the best overall performance, with the semantic encoding method achieving an F1 score of 91.07%, an improvement of 5.26% over original encoding, and accuracy increasing from 84.20% to 89.75%. The RoBERTa model also achieved significant improvements under the semantic encoding method, with accuracy improving by 5.60% and F1 score improving by 5.82%, reaching 90.67%.

Notably, all models achieved the greatest improvements in recall, with BERT, SciBERT, and RoBERTa improving by 6.10%, 7.80%, and 8.50% respectively, indicating that the semantic feature encoding mechanism has significant advantages in identifying positive samples. The improvements in precision were relatively stable, with the three models improving by 2.10%, 2.30%, and 2.70% respectively, demonstrating that this method effectively improves recall while maintaining high precision. This means that by replacing [CLS] tokens with topical semantic features and [SEP] tokens with structural semantic features, semantic information can be more deeply integrated into the attention computation process of PLMs. Compared to traditional feature concatenation or simple fusion methods, the replacement strategy in this paper enables semantic features to play roles at key positions: topical features at [CLS] positions can better aggregate global semantic representations, while structural features at [SEP] positions can more accurately model hierarchical structural information of texts. Furthermore, SciBERT's specialization in scientific text domains makes it perform more prominently when combined with semantic features, further proving the advantages of combining domain-adaptive pre-trained models with semantic feature encoding mechanisms.

6.2 Effectiveness Analysis of Binary Classification Models with Embedded Universal Semantic Features

To verify the practical application effectiveness of the proposed semantic feature encoding mechanism in scientific text value sentence recognition tasks, we conducted comparison experiments on the scientific literature value sentence dataset. This paper selected full texts from general domain scientific literature and constructed the corpus using manual annotation and iterative semi-automatic annotation methods. The dataset contains 23,912 scientific literature sentences with a 1:1 positive-to-negative sample ratio, divided into training set (19,130 samples), validation set (2,391 samples), and test set (2,391 samples) in an 8:1:1 ratio. The experiment designed 5 comparison methods covering different technical approaches including PLMs fine-tuning, LLMs zero-shot learning, semantic feature enhancement, and parameter-efficient fine-tuning to comprehensively evaluate the effectiveness of the proposed method. The PLMs fine-tuning parameters are the same as in Section 3.5.1; LLMs fine-tuning adopts the QLora (Dettmers et al., 2023) parameter-efficient fine-tuning method with the following parameter settings: rank=64, alpha=16.

Considering the input requirements of different types of models, this experiment adopted two different fine-tuning data formats for PLMs and LLMs:

(1) PLMs fine-tuning data format

For BERT-base, RoBERTa-base, and SciBERT, the standard classification task data format is adopted, containing Label and Sentence fields, where Label=0 indicates non-value sentences and Label=1 indicates value sentences. Specific data examples are shown in Table 5.

Table 5: PLMs fine-tuning data format.

Label	Sentence
0	Deep learning algorithms have been widely applied in various domains and achieved remarkable success.
1	This study aims to develop a novel neural architecture that can significantly improve the accuracy of text classification tasks while reducing computational complexity by 40%.

(2) Instruction fine-tuning dataset

For LLMs, the instruction fine-tuning format is adopted, containing Instruction, Input, and Output fields. The instruction section provides detailed descriptions of the definition and judgment criteria for value sentence recognition tasks, with the specific format as follows:

"Instruction": "Determine if the following sentence is a research value sentence. Research value sentences in scientific literature are sentences that explicitly describe the specific contributions, significance, or potential impact of the research work. They clearly state the research value, importance, or benefits that the study provides to the academic field or practical applications. Output 'True' if it is a research value sentence, and 'False' if it is not.",

"Input": "Our proposed method demonstrates superior performance on benchmark datasets, achieving state-of-the-art results with 15% improvement in accuracy compared to existing approaches.",

"Output": "True"

The experimental results are shown in Table 6. The results indicate that the proposed semantic feature encoding mechanism achieved consistent performance improvements across models with different architectures.

Table 6: The experimental results about instruction fine-tuning.

Methods	Models	A (%)	P (%)	R (%)	F1 (%)
Fine-tuning PLMs	BERT-base	84.32	83.15	85.62	84.37
	RoBERTa-base	86.45	85.73	87.28	86.50
	SciBERT	88.76	88.42	89.15	88.78
Base-LLMs	Qwen3-14B-base	79.23	76.84	82.47	79.55
	LLaMa4-17B-base	81.67	80.15	83.52	81.80
	GLM4-9B-base	77.89	75.62	80.73	78.10
LLMs Embedded Semantic Features	Qwen3-14B Embedded Semantic Features	86.75	85.92	87.84	86.87
	LLaMa4-17B Embedded Semantic Features	**89.34**	**88.67**	**90.15**	**89.40**
	GLM4-9B Embedded Semantic Features	84.56	83.74	85.62	84.67
QLora LLMs	Qwen3-14B-QLora	90.12	89.75	90.58	90.16
	LLaMa4-17B-QLora	91.83	91.46	92.27	91.86
	GLM4-9B-QLora	88.94	88.31	89.67	88.98
QLora LLMs Embedded Semantic Features	Qwen3-14B-QLora Embedded Semantic Features	**94.15**	**93.82**	**94.56**	**94.19**
	LLaMa4-17B-QLora Embedded Semantic Features	92.67	92.34	93.12	92.73
	GLM4-9B-QLora Embedded Semantic Features	91.28	90.85	91.84	91.34

(1) Independent effectiveness analysis of semantic feature encoding

In value sentence recognition tasks, incorporating semantic feature encoding significantly improves model recognition performance. Using LLaMa4-17B as the base model, the value sentence recognition accuracy after incorporating semantic features reached 89.34%, an improvement of 7.67% over the Base model without semantic features. Particularly, the recall rate reached 90.15%, an improvement of 6.63% over the Base model's recall rate, indicating that semantic feature encoding enables the model to capture the vast majority of value sentences, thereby significantly improving model recognition accuracy. The reason is that topical and structural semantic features specific to value sentences help the model further capture the linguistic patterns and semantic structures of value sentences. By directly embedding semantic representations into [CLS] and [SEP] positions, the model is assisted in focusing on semantic information most relevant to value sentence recognition, compensating for deep semantic associations that traditional methods might overlook.

(2) Collaborative analysis of instruction fine-tuning and semantic feature encoding

The collaborative mechanism of parameter-efficient fine-tuning and semantic feature encoding improved LLMs' performance in binary classification tasks. Using Qwen3-14B as the base model, the QLora fine-tuned version of LLaMa4-17B base model achieved an F1 score of 91.86%, with F1 score improving by 0.87% after combining semantic feature encoding; the GLM4-9B base model's F1 score improved from 88.98% to 91.34% after incorporating semantic features on the QLora fine-tuning basis. Particularly, Qwen3-14B-QLora + semantic features showed the best overall performance, achieving an accuracy of 94.19%, 2.55% higher than the QLora fine-tuned version. Semantic Feature Ablation Experiments

To thoroughly verify the specific contributions of topical semantic features and structural semantic features in the proposed semantic feature encoding mechanism, we designed systematic feature ablation experiments on the DBPedia dataset. This experiment analyzes the independent contributions and synergistic effects of each feature on model performance by progressively removing different semantic feature components. The feature ablation experiment designed the following 4 configuration schemes, using SciBERT as the base model for comparative analysis: (1) Original SciBERT: maintains traditional [CLS] and [SEP] tokens as the baseline method; (2) SciBERT + Topical Semantic Features: only replaces [CLS] tokens with topical semantic features; (3) SciBERT + Structural Semantic Features: only replaces [SEP] tokens with structural semantic features; (4) SciBERT + Complete Semantic Features: simultaneously embeds both topical and structural semantic features. All experiments used the same hyperparameter settings: learning rate of 1e-5, batch size of 32, 5 training epochs, and AdamW optimizer.

The results of semantic feature ablation experiments are shown in Table 7.

(1) Core contribution analysis of topical semantic features

From the perspective of independent effects of topical semantic features, the SciBERT model with only embedded topical semantic features achieved significant improvements across all evaluation metrics. The F1 score of SciBERT + topical semantic features reached 88.35%, an improvement of 2.54% over original SciBERT, with accuracy improving from 84.20% to 87.45%. Particularly noteworthy is that topical semantic features showed the most significant improvement in recall, from 81.60% to 85.20%, an increase of 3.60%, indicating that topical semantic features can effectively reduce false negative samples and improve the model's ability to recognize positive samples. The reason is that topical semantic features can more precisely capture the core semantic content of texts through dynamic window-based local-global feature extraction methods. After replacing the [CLS] position, the global semantic representation becomes more focused on the topical information of texts, thereby enhancing the model's ability to distinguish between different categories of texts.

Table 7: The results of semantic feature ablation experiments.

Models	A (%)	P (%)	R (%)	F1 (%)
SciBERT	84.20	90.50	81.60	85.81
SciBERT + Topical Semantic Features	87.45	91.75	85.20	88.35
SciBERT + Structural Semantic Features	86.10	91.20	83.80	87.35
SciBERT+ Complete Semantic Features	**89.75**	**92.80**	**89.40**	**91.07**

(2) Auxiliary enhancement analysis of structural semantic features

From the perspective of independent contribution of structural semantic features, embedding structural semantic features alone also brought obvious performance improvement effects. The F1 score of SciBERT + structural semantic features was 87.35%, an improvement of 1.54% over original SciBERT, with accuracy improving by 1.90%. Compared to topical semantic features, the improvement magnitude of structural semantic features was relatively smaller, but it showed stable performance in precision, improving from 90.50% to 91.20%, an increase of 0.70%.

(3) Synergistic effect analysis of complete semantic features

From the perspective of synergistic effects between topical and structural semantic features, the complete semantic feature configuration achieved the best comprehensive performance. The F1 score of SciBERT + complete semantic features reached 91.07%, an improvement of 5.26% over original SciBERT. This improvement magnitude exceeded the simple additive effect of using topical semantic features alone (2.54% improvement) and structural semantic features alone (1.54% improvement), indicating significant synergistic enhancement mechanisms between the two semantic features. This means that topical and structural semantic features form effective complementarity in function: topical semantic features focus on capturing content semantics of texts, while structural semantic features concentrate on modeling organizational forms of texts. Their combination can provide more comprehensive semantic understanding capabilities for the model.

7 CONCLUSIONS

This paper proposes a LLM text classification method with embedded semantic feature encoding that constructs a dual semantic feature encoding mechanism, embedding topical semantic features and structural semantic features into the [CLS] and [SEP] positions of LLMs, respectively, thereby achieving deep fusion between semantic features and internal model representations. The method employs dynamic window-based local-global feature extraction strategies to capture topical semantic features, utilizes hierarchical structural aggregation mechanisms to capture organizational semantic information of texts, and directly embeds semantic information into critical positions through feature replacement encoding strategies, enhancing the model's understanding of holistic textual semantics.

Experimental results demonstrate that the proposed semantic feature encoding mechanism achieves significant performance improvements across multiple benchmark datasets. On the DBPedia dataset, the semantically encoded SciBERT model achieves an F1-score of 91.07%, representing a 5.26% improvement over the original encoding approach, with accuracy increasing from 84.20% to 89.75%. In the scientific literature value sentence identification task, Qwen3-14B combined with QLora fine-tuning and semantic feature encoding achieves an F1-score of 94.19%, showing a 14.64% improvement over the baseline model, validating the effectiveness of the semantic feature encoding mechanism. Ablation experiments further confirm the synergistic effects of topical semantic features and structural semantic features, with the complete semantic feature configuration achieving performance improvements that exceed the simple additive effects of individual features, indicating that the two types of semantic features form effective functional complementarity. Compared to traditional feature concatenation or simple fusion approaches, our feature replacement encoding strategy enables semantic information to function at critical positions in attention computation, achieving deep integration between semantic features and the internal mechanisms of LLMs.

ACKNOWLEDGEMENTS

This study was funded by the National Key R&D Program of China(2022YFF0711900)

REFERENCES

Feng, P., Zhang, X., Zhao, J., Wang, Y., & Huang, B. (2023). Relation Extraction Based on Prompt Information and Feature Reuse. *Data Intelligence*, 5(3), 824-840.

Li, Y., Zhang, M., Zhang, Z., et al. (2024). Decoding the Essence of Scientific Knowledge Entity Extraction: An Innovative MRC Framework with Semantic Contrastive Learning and Boundary Perception. *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 1-12.

Wang, M., Kim, J., & Yan, Y. (2025). Syntactic-Aware Text Classification Method Embedding the Weight Vectors of Feature Words. *IEEE Access*, 13, 37572-37590.

Liu, C., Zhang, H., Zhao, K., et al. (2024). LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification. *arXiv preprint* arXiv:2406.03725.

Devlin, J., Chang, M. W., Lee, K., et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171-4186.

Ezugwu, A. E., Ho, Y.-S., Egwuche, O. S., et al. (2024). Classical Machine Learning: Seventy Years of Algorithmic Learning Evolution. *Data Intelligence*. https://www.sciengine.com/doi/10.3724/2096-7004.di.2024.0051

Zhao, J., Lan, M., Niu, Z. Y., et al. (2015). Integrating word embeddings and traditional NLP features to measure textual entailment and semantic relatedness of sentence pairs. *Proceedings of the International Joint Conference on Neural Networks*, 1-7.

Pangakis, N., & Wolken, S. (2024). Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels. *arXiv preprint* arXiv:2406.17633.

Kokkodis, M., Demsyn-Jones, R., & Raghavan, V. (2025). Beyond the Hype: Embeddings vs. Prompting for Multiclass Classification Tasks. *arXiv preprint* arXiv:2504.04277.

Wang, M., Zhang, Z., Li, H., et al. (2024). An Improved Meta-Knowledge Prompt Engineering Approach for Generating Research Questions in Scientific Literature. *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR*, 457-464.

Achiam, J. (2023). GPT-4 technical report. *arXiv preprint* arXiv:2303.08774.

Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971.

Guo, Z., Jiao, K., Yao, X., et al. (2024). USTC-BUPT at SemEval-2024 Task 8: Enhancing Machine-Generated Text Detection via Domain Adversarial Neural Networks and LLM Embeddings. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 1511-1522..

Li X, Zhang Z, Liu Y et al. (2022). A Study on the Method of Identifying Research Question Sentences in Scientific Articles[J]. *Library and Information Service*, 67(9) pp 132-140

Liu P, Cao Y (2022). A named entity recognition method for Chinese winter sports news based on RoBERTa-WWM[C] *Proc. 3rd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*. pp 785-790.

Liu X, Zhao W, Ma H (2022). Research on domain-specific knowledge graph based on the RoBERTa-wwm-ext pretraining model[J]. *Comput. Intell. Neurosci*., pp 1-11.

Han Y (2023). Advancing Text Analytics: Instruction Fine-Tuning of QianWen-7B for Sentiment Classification[C] *Proceedings of the 2023 4th International Conference on Big Data Economy and Information Management*. pp 90-93.

Zhang Y, Wang M, Ren C, et al. (2024). Pushing the limit of LLM capacity for text classification[EB/OL]. *arXiv*:2402.07470.

Chae Y, Davidson T (2024). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning[J]. *Sociological Methods & Research*, 00491241251325243.

[21] Fatemi S, Hu Y, Mousavi M (2025). A Comparative Analysis of Instruction Fine-Tuning Large Language Models for Financial Text Classification[J]. *Management Information Systems*, 16(1).

Peng L, Shang J (2024). Incubating text classifiers following user instruction with nothing but LLM[EB/OL]. *arXiv*:2404.10877.

Meguellati E, Zeghina A, Sadiq S, et al. (2025). LLM-based Semantic Augmentation for Harmful Content Detection[EB/OL]. *arXiv*:2504.15548.

Guo Y, Ovadje A, Al-Garadi M A, et al. (2024). Evaluating large language models for health-related text classification tasks with public social media data[J]. *Journal of the American Medical Informatics Association,* 31(10) pp 2181-2189.

Lehmann J, Isele R, Jakob M, et al. (2015). DBpedia--a large-scale, multilingual knowledge base extracted from Wikipedia[J]. *Semantic Web,* 6(2): 167-195.

Beltagy I, Lo K, Cohan A (2019). SciBERT: A pretrained language model for scientific text[EB/OL]. *arXiv*:1903.10676.

Yang A, Li A, Yang B, et al. (2025). Qwen3 technical report[EB/OL]. [2025]. *arXiv*:2505.09388.

Meta A I. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation[EB/OL]. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

GLM T, Zeng A, Xu B, et al. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools[EB/OL]. *arXiv*:2406.12793.

Loshchilov I, Hutter F (2019). Decoupled weight decay regularization[C] *International Conference on Learning Representations (ICLR)*.

Dettmers T, Pagnoni A, Holtzman A, et al. (2023). QLoRA: Efficient finetuning of quantized LLMs[C] *Advances in Neural Information Processing Systems (NeurIPS)*.