Innovative Sentence Classification in Scientific Literature: A Two-Phase Approach with Time Mixing Attention and Mixture of Experts

Meng Wang[1][a], Mengting Zhang[1,2][b], Hanyu Li[1,2][c],
Jing Xie[1][d], Zhixiong Zhang[1,2][e], Yang Li[1,2][f] and Gaihong Yu[1][g]
[1]*National Science Library, Chinese Academy of Sciences, Beijing 100190, China*
[2]*School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China*

Keywords: Innovative Sentence Identification, Multi-Class Text Classification, Time Mixing Attention, Mixture of Experts, Generative Semantic Data Augmentation.

Abstract: Accurately classifying innovative sentences in scientific literature is essential for understanding research contributions. This paper proposes a two-phase classification framework that integrates a Time Mixing Attention (TMA) mechanism and a Mixture of Experts (MoE) system to enhance multi-class innovation classification. In the first phase, TMA improves long-range dependency modeling through temporal shift padding and sequence slice reorganization. The second phase employs an MoE-based approach to classify theoretical, methodological, and applied innovations. To mitigate class imbalance, a generative semantic data augmentation method is introduced, improving model performance across different innovation categories. Experimental results demonstrate that the proposed two-phase SciBERT+TMA model achieves the highest performance, with a macro-averaged F1-score of 90.8%, including 95.1% for theoretical innovation, 90.8% for methodological innovation, and 86.6% for applied innovation. Compared to the one-phase SciBERT+TMA model, the two-phase approach significantly improves precision and recall, highlighting the benefits of progressive classification refinement. In contrast, the best-performing LLM baseline, Ministral-8B-Instruct, achieves a macro-averaged F1-score of 85.2%, demonstrating the limitations of prompt-based inference in structured classification tasks. The results underscore the advantage of a domain-adapted approach in capturing fine-grained distinctions in innovation classification. The proposed framework provides a scalable solution for multi-class sentence classification and can be extended to broader academic classification tasks. Model weights and details are available at *https://huggingface.co/wmsr22/Research_Value_Generation/tree/main*.

1 INTRODUCTION

Text classification, a fundamental natural language processing (NLP) task, involves categorizing textual units—ranging from documents to sentences—into predefined categories based on their meaning or function. (You et al., 2019). This task plays a critical role in the processing of scientific literature, as it facilitates the capture and analysis of complex semantic structures, the understanding of intricate linguistic patterns, and the extraction of key information (Wang et al., 2025). Consequently, optimizing and improving text classification models, particularly in the context of scientific literature, has become one of the central research topics in the field of NLP (Shang et al., 2024).

The self-attention mechanism, as a core component of state-of-the-art neural architectures in text classification, allows the model to learn the internal structure of input data and focus on the most relevant parts during information processing. This capability has proven to be highly effective in a variety of text classification tasks (Guo et al., 2020). However, with the rapid growth of scientific literature and the acceleration of interdisciplinary research, the variety of sentence types within the literature has increased, often exhibiting significant class imbalance. On one hand, certain sentence types are relatively rare and dispersed throughout the text, which places higher demands on the model's ability to capture long-distance

[a] https://orcid.org/0009-0009-7780-6516
[b] https://orcid.org/0000-0002-9941-7548
[c] https://orcid.org/0000-0003-1426-3242
[d] https://orcid.org/0000-0001-6698-1786
[e] https://orcid.org/0000-0003-1596-7487
[f] https://orcid.org/0009-0008-8451-1452
[g] https://orcid.org/0000-0003-1301-2871

Wang, M., Zhang, M., Li, H., Xie, J., Zhang, Z., Li, Y., Yu and G.
Innovative Sentence Classification in Scientific Literature: A Two-Phase Approach with Time Mixing Attention and Mixture of Experts.
DOI: 10.5220/0013513200003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 382-389
ISBN: 978-989-758-758-0; ISSN: 2184-285X
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

dependencies. On the other hand, even within the same broad category, there is considerable variation in the number of sentences across different subtypes, further exacerbating the class imbalance and reducing the accuracy of multi-class classification models (Oida-Onesa and Ballera, 2024). Therefore, integrating temporal information into the self-attention mechanism and optimizing the overall classification architecture are critical for improving performance in these highly imbalanced and complex tasks.

To address these issues, this paper proposes a two-phase classification framework based on an improved self-attention mechanism that incorporates time-mixing information to enhance the model's capability in handling long-range dependencies. Additionally, by designing a hierarchical classification architecture and incorporating a generative semantic-based data augmentation method, we aim to improve the accuracy and generalization ability of the multi-classification model.

The main contributions of this paper are summarized as follows:

- To improve the accuracy of multi-classification models, a two-phase classification architecture is designed, which integrates the Time Mixing Attention (TMA) mechanism and a mixture of experts (MoE) system. This architecture enhances the model's ability to recognize different types of sentences through hierarchical feature extraction and classification decisions.
- To capture long-range dependencies and contextual semantic information, the self-attention mechanism is introduced into the model used in the first phase, with dynamic temporal encoding of context, thereby improving the model's ability to capture sequential dynamic behaviors.
- To address the issue of data imbalance, a generative semantic-based data augmentation method is proposed. This method expands the training samples of rare categories by generating data that maintain semantic consistency, thereby improving the model's classification performance across all categories.
- To validate the effectiveness of the proposed method, comparative experiments are conducted on innovative sentences in scientific literature, including both pre-trained language models (PLMs) and large language models (LLMs). The experimental results demonstrate significant improvements on macro-average metrics.

2 RELATED WORK

In this paper, we conduct a literature review on multi-class text classification, the self-attention mechanism, and data augmentation approaches, as these areas provide the essential theoretical and technical foundation for the development of efficient text classification models and the mitigation of data imbalance challenges.

2.1 Multi-Class Text Classification

Multi-class text classification is the task of categorizing a given text into one of several predefined categories, with each text belonging to only one class from a set of possible labels (Wang et al., 2024). Given the importance of this task, numerous approaches have been proposed to improve classification accuracy and efficiency. Jain et al. (Jain et al., 2024) developed a hierarchical text classification framework that encodes dynamic text representations using language models and introduces a horizontal guidance loss function to capture relationships between text and label semantics, thus adapting language models to domain knowledge. Afzal et al. (Afzal et al., 2024) proposed a Transformer-based active learning approach for multi-class text annotation and classification, which utilizes deep learning techniques to enhance annotation efficiency and improve classification performance, particularly for unstructured medical data. Similarly, Le et al. (Le et al., 2024) introduced CoLAL, an active learning algorithm that combines noisy labels and predictions from the primary model to select diverse and representative samples, significantly enhancing performance in text-based active learning.

2.2 Self-Attention Mechanism

To address long-range dependency issues, researchers have primarily focused on enhancing the self-attention mechanism, particularly by improving its ability to capture relationships across positions in a sequence. Vaswani et al. (Vaswani, 2017) introduced positional embeddings and integrated the attention mechanism into the Transformer architecture, thus obviating the need for recurrence and convolution, while effectively capturing long-range dependencies and contextual relationships. Yu et al. (Yu et al., 2024) proposed a dual attention module designed to capture dependencies between different sequences and variations within individual sequences, using a learnable decomposition strategy for multivariate time series prediction, which improved the

capture of dynamic trend information for more accurate time series forecasting.

Data augmentation is a method to generate additional data by manipulating original samples to increase diversity and quantity while preserving their core characteristics (Bayer et al., 2022). In the context of text, data augmentation involves creating new training samples through semantic-preserving transformations, rule-based replacements, or generative models, thereby expanding the scale and diversity of datasets to improve model generalization and robustness.

3 OVERVIEW

This section introduces the task of identifying and classifying innovative sentences in scientific literature. The primary objective of this task is to categorize sentences into two main groups: innovative and non-innovative. Additionally, for sentences identified as innovative, the task further classifies them into three distinct categories: theoretical innovation, methodological innovation, and applied innovation. The task is structured into two layers, each employing specialized techniques to address different aspects of the classification problem, ensuring a more precise and robust categorization process.

3.1 Task Objectives

The task involves two main objectives: the identification of innovative sentences and their subsequent classification into specific categories.

3.1.1 Innovative Sentence Identification

The first objective is to identify innovative sentences. This step focuses on detecting sentences that introduce new ideas, theories, or advancements, which are classified as innovative sentences. In contrast, non-innovative sentences either restate established knowledge or provide general descriptions without offering novel insights or perspectives.

3.1.2 Innovation Classification

Once innovative sentences are identified, the next step is to classify them into one of three types of innovation:

- **Theoretical Innovation:** Sentences that propose new theories or frameworks, offering fresh perspectives on existing problems and advancing theoretical understanding.
- **Methodological Innovation:** Sentences that introduce novel research methods, techniques, or tools, enhancing the way research is conducted through improved experimental designs or data analysis approaches.
- **Applied Innovation:** Sentences that demonstrate the practical application of existing theories or methodologies in new contexts, addressing real-world problems or societal needs.

3.2 Approach Design

3.2.1 First Phase: Innovative Sentence Identification

The first phase is dedicated to identifying innovative sentences within a given text. To achieve this, we begin by utilizing a pre-trained model to encode each token in the sentence into high-dimensional embedding vectors. These embeddings capture rich semantic information and contextual dependencies for each token within the sentence. Following the encoding step, the **Time Mixing Attention (TMA)** mechanism, as detailed in Section 4.1, is applied to model inter-token dependencies and capture the contextual relationships between the token representations. Subsequently, the resulting feature representations are passed through a fully connected layer, which maps them into a two-dimensional space. This projection enables the final classification of sentences into two categories: innovative and non-innovative, based on the aggregated representation of the sentence.

3.2.2 Second Phase: Innovation Classification

The second phase involves classifying the identified innovative sentences into one of three categories: theoretical, methodological, or applied innovation. This classification process is facilitated through the integration of **Mixture of Experts (MoE)**, as outlined in Section 4.2, and **Generative Semantic Data Augmentation**, discussed in Section 4.3. The MoE system enables the dynamic selection of specialized experts based on the content of each sentence, ensuring precise categorization. Simultaneously, the generative augmentation techniques leverage large language models to address issues of class imbalance and data scarcity by enriching the training dataset with diverse sentence variations, thereby increasing the model's ability to handle a wide range of sentence structures and enhancing its overall robustness.

4 METHODOLOGY

4.1 Time Mixing Attention Mechanism

We propose the Time Mixing Attention (TMA) mechanism, which integrates time-shift padding, embedding dimension slicing and recombination, and self-attention to effectively capture both local subspace features and global temporal dependencies.

4.1.1 Time-Shift Padding

To provide contextual support for boundary positions in time series data, TMA employs time-shift padding as a foundational step. In time series sequences, the initial and final positions often lack sufficient contextual information due to the absence of preceding or succeeding elements. Time-shift padding addresses this by symmetrically appending zero vectors to both ends of the embedding sequence, providing additional contextual "buffers".

Let the input sequence be denoted as $S = \{s_1, \ldots, s_i, \ldots, s_T\}$, and its corresponding embedding matrix, generated by a pre-trained language model, as $X = [x_1, \ldots, x_i, \ldots, x_T]$, where $X \in \mathbb{R}^{T \times d}$, and x_i represents the d-dimensional embedding vector of the i-th element of S. Zero-padding along the temporal dimension produces the extended embedding matrix $X' = \{x_0, x_1, \ldots, x_i, \ldots, x_T, x_{T+1}\}$, where x_0 and x_{T+1} are zero vectors acting as padding for the start and end positions respectively, as illustrated in Figure 1. This operation ensures sufficient contextual support for boundary positions while maintaining the integrity of the original sequence data by using zero-padding, which introduces no additional semantic or structural information and preserves the neutrality of the original embeddings.

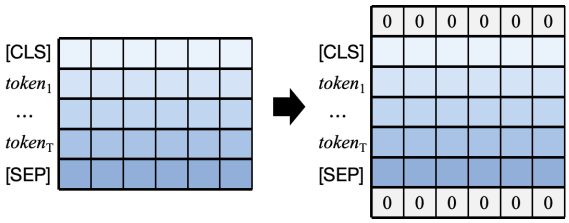

Figure 1: Time-shift Padding.

4.1.2 Embedding Dimension Slicing and Recombination

After applying time-shift padding, the extended embedding matrix $X' \in \mathbb{R}^{(T+2) \times d}$, where each timestep $x_i \in \mathbb{R}^d$ (with $i \in \{1, 2, ..., T+1\}$) is a d-dimensional embedding vector, undergoes embedding dimension slicing and recombination to capture local subspace information. Each embedding vector x_i is partitioned into three equal subspaces along its embedding dimension (d is typically 768), as follows:

- The starting part $x_{i,\text{start}} = x_i[: d/3]$, corresponding to the first third of the embedding vector;
- The middle part $x_{i,\text{middle}} = x_i[d/3 : 2d/3]$, corresponding to the middle third of the embedding vector;
- The ending part $x_{i,\text{end}} = x_i[2d/3 :]$, corresponding to the last third of the embedding vector.

These slices are then recombined in a shifted manner to form the final vector. Specifically, the slices from consecutive timesteps are shifted as shown in Figure 2, and the formula is as follows:

$$x'_i = \text{Concat}(x_{i-1,\text{start}}, x_{i,\text{middle}}, x_{i+1,\text{end}}) \quad (1)$$

where $i \in \{1, 2, ..., T\}$. This operation ensures that each timestep's embedding vector x'_i incorporates information from neighboring time spans. The starting part $x_{i-1,\text{start}}$ comes from the previous timestep, the middle part $x_{i,\text{middle}}$ comes from the current timestep, and the ending part $x_{i+1,\text{end}}$ comes from the subsequent timestep.

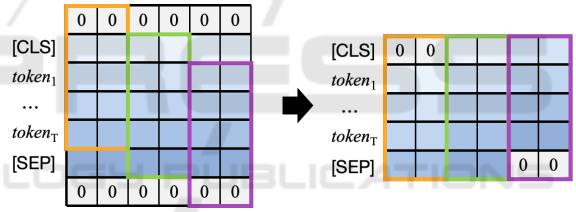

Figure 2: Embedding Dimension Slicing and Recombination.

By shifting the slices in this manner, the recombination process captures the temporal dependencies between consecutive timesteps, allowing the model to integrate information from neighboring time intervals. The result is a new embedding matrix $X'' \in \mathbb{R}^{T \times d}$, where each timestep's embedding vector x'_i now reflects a broader context by incorporating local and adjacent temporal information. This approach helps the model better capture both short- and long-range dependencies within the time series data.

4.1.3 Self-Attention Mechanism

Building upon the embedding matrix obtained from the previous step, the self-attention mechanism is employed to model the dynamic relationships between timesteps. First, the embedding matrix X'' is projected into three matrices: Query (Q), Key (K), and Value (V), as follows:

$$Q = X''W^Q, \quad K = X''W^K, \quad V = X''W^V \quad (2)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_m}$ are learned projection matrices. Next, the attention weights $\mu_{i,j}$ between timesteps are computed using the vectors derived from the matrices Q and K. Specifically, the attention score between timesteps i and j is given by:

$$\omega_{i,j} = \frac{q_i k_j^T}{\sqrt{d_m}} \tag{3}$$

where q_i and k_j are the i-th and j-th rows of the matrices Q and K, respectively. The attention weight $\mu_{i,j}$ is then computed using the softmax function as follows:

$$\mu_{i,j} = \frac{\exp(\omega_{i,j})}{\sum_{j=1}^{T} \exp(\omega_{i,j})} \tag{4}$$

These attention weights $\mu_{i,j}$ reflect the degree of relevance between timestep x'_i and timestep x'_j, with larger values indicating stronger temporal dependencies.

Finally, the weighted sum of the Value vectors is computed to generate the new representation, as follows:

$$z_i = \sum_{j=1}^{T} \mu_{i,j} v_j \tag{5}$$

where v_j is the j-th row of the Value matrix V, and z_i represents the final feature vector for timestep i. This vector integrates information from all timesteps, weighted by their respective attention scores $\mu_{i,j}$. This mechanism allows the model to dynamically focus on relevant parts of the sequence, effectively capturing both short- and long-range temporal dependencies.

4.2 Mixture of Experts

In this paper, we construct independent expert models for each type of innovative sentence: theoretical innovation, methodological innovation, and applied innovation. These expert models are the same ones used in the innovative sentence identification task, which incorporates the Time Mixing Attention (TMA) mechanism. To ensure that each expert model effectively focuses on its designated innovation dimension, we create three distinct annotated datasets, each corresponding to one of the innovation types (theoretical, methodological, or applied innovation). Specifically, the annotated datasets are denoted as $(x_i^{(j)}, y_i^{(j)})_{i=1}^{N_j}$ with $j \in \{1,2,3\}$, where $x_i^{(j)}$ represents the input texts, $y_i^{(j)}$ represents the corresponding true labels, and N_j is the total number of samples for the j-th innovation type.

The training objective for each expert model is to minimize the binary cross-entropy loss (BCE) between the model's predicted output and the true label. The objective function for the j-th expert model is defined as:

$$\mathcal{L}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathrm{BCE}(f_j(x_i^{(j)}), y_i^{(j)}) \tag{6}$$

where $f_j(x_i^{(j)})$ denotes the output of the j-th expert model for the input $x_i^{(j)}$.

After training the expert models, we perform inference by selecting the most appropriate expert for each input sentence. During inference, a hard routing mechanism is employed, which is based on the probability distribution computed as follows:

$$p_j = \sigma(f_j(x)) \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $f_j(x)$ is the output of the j-th expert model for the input x. The probability p_j represents the likelihood that the j-th expert is the most suitable for making the prediction for the given input. The expert model with the highest probability is selected for final inference, as indicated by:

$$j_{\max} = \arg\max_j p_j \tag{8}$$

This hard routing mechanism ensures that only the most appropriate expert contributes to the final prediction, thereby reducing computational complexity and improving the model's efficiency. The final prediction y_{final} is made using the selected expert model:

$$y_{\text{final}} = f_{j_{\max}}(x) \tag{9}$$

4.3 Generative Semantic Data Augmentation

In the fine-grained Innovation Classification task, class imbalance and data scarcity pose significant challenges across the three categories of innovative sentences. To mitigate these issues, this study proposes a generative semantic data augmentation approach leveraging large language models, which aims to enhance the diversity and quality of training data, thereby improving the model's classification performance.

In the data augmentation process, we begin by combining labeled innovative sentences with specific prompts to create the instructions required for generating sentences using the large language model. These instructions consist of three main components: task description, seed sentences, and generation requirements. The prompt we constructed is as follows:

Generate semantically related sentences for each seed sentence, with varied expressions, to enrich the dataset: Seed 1: [We designed a novel experimental

Table 1: Experimental Results for Innovative Sentence Identification.

Model	Acc%	P%	R%	F1%
BERT	83.2	82.0	**90.8**	86.1
RoBERTa	83.9	84.9	87.6	86.2
SciBERT	84.3	83.4	**90.8**	86.9
BERT + TMA	83.3	85.5	85.5	85.5
RoBERTa + TMA	83.8	83.0	90.4	86.6
SciBERT + TMA	**85.3**	**86.1**	88.8	**87.4**

approach combining flow cytometry and mass spectrometry to analyze cellular responses under different conditions.] Seed 2: [The developed system has immediate applications in remote patient monitoring, enabling real-time health status tracking in rural areas.]. Requirements: 1) Semantically related to seeds, reflecting methodological innovation and application innovation. 2) Follow application scenario description style and norms. 3) Vary in expression, similar in length and complexity. 4) Demonstrate practical implementation and impact. 5) Generate sentences in English, enclosed in square brackets [].

5 EXPERIMENTS

The experiments consist of two parts: the first part focuses on identifying innovative sentences, where the model is trained to recognize and extract innovation-related sentences from scientific literature. The second part addresses innovation classification, where the identified sentences are categorized into three distinct types: theoretical, methodological, and applied innovations.

5.1 Experiment for Innovative Sentence Identification

The objective of this experiment was to validate the effectiveness of the Time Mixing Attention (TMA) mechanism in identifying innovative sentences from scientific literature. To construct the dataset, sentences were extracted from relevant sections, resulting in a total of 23,912 annotated sentences, with 10,036 labeled as non-innovative sentences and 13,876 as innovative sentences. The dataset was split into training, validation, and test sets in an 8:1:1 ratio while ensuring balanced distributions across subsets. The experiment utilized multiple models, including BERT (Devlin, 2018), RoBERTa (Liu, 2019), and SciBERT (Beltagy et al., 2019), both with and without the integration of TMA, to assess its impact on classifying innovation and non-innovative sentences based on their semantic content. The training configuration employed a learning rate of 1×10^{-5}, a batch size of 5, and a single training epoch. Performance evaluation was conducted using macro-average precision (P), recall (R), F1-score (F1) and accuracy (Acc) to comprehensively assess the effectiveness of the models and the impact of TMA.

As shown in Table 1, the integration of the TMA mechanism led to significant improvements in model performance. SciBERT with TMA achieved the best results, with a test accuracy of 85.3% and an F1-score of 87.4%, outperforming all other models. In addition, All models incorporating TMA achieved an average increase of 0.1% in F1-score compared to their counterparts without the mechanism.

5.2 Experiment for Innovation Classification

The goal of this experiment was to evaluate the effectiveness of our proposed two-phase innovation classification approach in distinguishing between different types of innovative sentences in scientific literature. The dataset was initially constructed through manual annotation, yielding a total of 13,876 labeled innovative sentences, categorized into three innovation types: 11,353 theoretical innovative sentences, 1,171 methodological innovative sentences, and 1,352 applied innovative sentences. Given the significant class imbalance, where methodological and applied innovative sentences were substantially fewer in number, we applied our proposed LLM-based generative semantic data augmentation to expand these categories. Specifically, we utilized Claude-3.5-Sonnet to generate additional innovative sentences while preserving linguistic diversity and semantic consistency. For theoretical innovative sentences, we selected 6,000 high-quality instances from the manually labeled data. For methodological innovative sentences, we expanded the dataset to 5,500 sentences, combining the original annotations with generated samples. Similarly, the applied innovation category was augmented to 5,000 sentences using a mix of original and generated

Table 2: Experimental Results for Two-phase Innovation Classification.

Model	Theoretical				Methodological				Applied				Macro avg			
	Acc%	P%	R%	F1%	Acc%	P%	R%	F1%	Acc%	P%	R%	F1%	Acc%	P%	R%	F1%
Ministral-8B-Instruct	86.7	86.3	86.9	86.6	85.2	84.8	85.5	85.1	84.1	83.8	84.3	84.0	85.3	85.0	85.6	85.2
LLAMA 3.1 8B-instruct	84.3	83.8	83.1	83.4	85.2	85.9	84.7	85.3	83.6	82.9	84.1	83.5	84.4	84.2	84.0	84.1
Phi 4-14B	85.8	86.2	85.5	85.8	84.5	84.2	84.7	84.4	83.2	83.5	83.0	83.2	84.5	84.6	84.4	84.5
Qwen2.5 -14B	85.6	84.9	85.3	85.1	84.2	83.8	84.5	84.1	83.1	82.8	83.3	83.0	84.3	83.8	84.4	84.1
SciBERT+TMA (one-phase)	91.2	96.1	85.2	90.3	91.2	90.4	**90.3**	90.4	**91.2**	83.7	74.6	78.9	**91.2**	90.1	83.4	86.5
SciBERT+TMA (two-phase)	**95.2**	**96.7**	**93.6**	**95.1**	**91.5**	**98.3**	84.3	**90.8**	86.3	**85.0**	**88.3**	**86.6**	91.0	**93.3**	**88.7**	**90.8**

data. To balance the dataset and reduce model bias, an equal number of non-innovative sentences was incorporated for each category, maintaining a 1:1 positive-to-negative sample ratio. The dataset was then stratified into training, validation, and test sets using an 8:1:1 split, ensuring a consistent distribution of positive and negative samples across all subsets.

The experiment evaluated multiple mainstream LLMs, including Ministral-8B-Instruct (Paramanayakam et al., 2024), LLAMA 3.1 8B-Instruct (Dubey et al., 2024), Phi 4-14B (Abdin et al., 2024), and Qwen2.5-14B (Yang et al., 2024), using a prompt-based classification approach, where sentences were categorized based on the following prompt:

Your task is to classify each sentence in a given text into one of four categories:

1.Theoretical Sentences: Sentences that discuss theoretical frameworks, conceptual definitions, theoretical models, or hypotheses

2.Methodological Sentences: Sentences that describe research methods, data collection processes, experimental designs, or analytical techniques

3.Applied Sentences: Sentences that discuss practical applications, implications, solutions, or recommendations

4.Other Sentences: Sentences that don't fit into the above categories

The following is an example:

Input: *"Recent advances in cognitive psychology have suggested that working memory capacity is not fixed but can be enhanced through training. To test this hypothesis, we recruited 150 undergraduate students and randomly assigned them to experimental and control groups. The experimental group underwent an 8-week computerized working memory training program, while the control group played casual computer games. Results showed that students who completed the training demonstrated significant improvements in both working memory tasks and academic performance, suggesting that such interventions could be valuable for educational programs."*

Output: *{ "sentences": [{ "text": "Recent advances in cognitive psychology have suggested that working memory capacity is not fixed but can be enhanced through training.", "classification": "theoretical" }, { "text": "To test this hypothesis, we recruited 150 undergraduate students and randomly assigned them to experimental and control groups.", "classification": "methodological" }, { "text": "The experimental group underwent an 8-week computerized working memory training program, while the control group played casual computer games.", "classification": "methodological" }, { "text": "Results showed that students who completed the training demonstrated significant improvements in both working memory tasks and academic performance, suggesting that such interventions could be valuable for educational programs.", "classification": "applied" }]*

Now, please extract the information from the following text:

Input:

In addition, we conducted experiments using SciBERT+TMA, the best-performing model from our first phase of innovative sentence identification. The classification was performed under two settings: (1) a one-phase classification approach, where sentences were directly categorized into non-innovation, theoretical innovation, methodological innovation, and applied innovation; and (2) our proposed two-phase model, which first identified innovative sentences and then further classified them into subcategories using our MoE-based approach.

As shown in Table 2, the proposed two-phase SciBERT+TMA model achieved the highest performance, with a macro-averaged F1-score of 90.8%. Specifically, it attained F1-scores of 95.1% for theoretical innovation, 90.8% for methodological innovation, and 86.6% for applied innovation. Compared to the one-phase SciBERT+TMA model, the two-phase approach demonstrated notable improvements, particularly in precision and recall, highlighting the benefits of progressive classification refinement.

6 CONCLUSIONS

This paper presents a two-phase classification framework for identifying innovative sentences in scientific literature, integrating a Time Mixing Attention (TMA) mechanism and a Mixture of Experts (MoE) model. The first phase enhances long-range dependency modeling using TMA, while the second phase employs MoE to classify sentences into theoretical, methodological, and applied innovation categories. Additionally, a generative semantic data augmentation method is introduced to address class imbalance and improve model performance.

Experimental results demonstrate that the proposed two-phase SciBERT+TMA model achieves superior performance, with a macro-averaged F1-score of 90.8%, outperforming the one-phase approach and all LLM baselines. Specifically, the F1-scores for theoretical, methodological, and applied innovation categories reach 95.1%, 90.8%, and 86.6%, respectively, highlighting the effectiveness of progressive classification refinement. Compared to direct classification, the MoE-based approach significantly improves precision and recall. Among LLMs, Ministral-8B achieves the best performance in prompt-based classification, with a macro-averaged F1-score of 85.2%, reinforcing the advantages of a domain-adapted framework over general-purpose LLM inference.

ACKNOWLEDGEMENTS

This study was funded by National Social Science Foundation of China (Grant No. 21&ZD329).

REFERENCES

Abdin, M., Aneja, J., Awadalla, H., and et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint*, arXiv:2404.14219.

Afzal, M., Hussain, J., Abbas, A., and et al. (2024). Transformer-based active learning for multi-class text annotation and classification. *Digital Health*, 10:20552076241287357.

Bayer, M., Kaufhold, M. A., and Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubey, A., Jauhri, A., Pandey, A., and et al. (2024). The llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783.

Guo, Q., Qiu, X., Liu, P., and et al. (2020). Multi-scale self-attention for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7847–7854.

Jain, V., Rungta, M., Zhuang, Y., and et al. (2024). Higen: Hierarchy-aware sequence generation for hierarchical text classification. *arXiv preprint arXiv:2402.01696*.

Le, L., Zhao, G., Zhang, X., and et al. (2024). Colal: Co-learning active learning for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13337–13345.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Oida-Onesa, R. and Ballera, M. A. (2024). Fine tuning language models: A tale of two low-resource languages. *Data Intelligence*, 6(4):946–967.

Paramanayakam, V., Karatzas, A., and Anagnostopoulos, I, e. a. (2024). Less is more: Optimizing function calling for llm execution on edge devices. *arXiv preprint*, arXiv:2411.15399.

Shang, S., Jiang, R., Shibasaki, R., and Yan, R. (2024). Foundation models for information retrieval and knowledge processing. *Data Intelligence*, 6(4):891–892.

Vaswani, A. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Wang, M., Kim, J., and Yan, Y. (2025). Syntactic-aware text classification method embedding the weight vectors of feature words. *IEEE Access*, 13:37572–37590.

Wang, M., Zhang, Z., Li, H., and Zhang, G. (2024). An improved meta-knowledge prompt engineering approach for generating research questions in scientific literature. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)*, volume 1, pages 457–464.

Yang, A., Yang, B., Zhang, B., and et al. (2024). Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115.

You, R., Zhang, Z., Dai, S., and et al. (2019). Haxmlnet: Hierarchical attention network for extreme multi-label text classification. *arXiv preprint arXiv:1904.12578*.

Yu, G., Zou, J., Hu, X., and et al. (2024). Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. *arXiv preprint arXiv:2402.12694*.