Transactions on Computational and Scientific Methods | Vo. 4, No. 3, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Multi-Label Text Classification with Large Language Models via Hierarchical Prompt Learning and Label Dependency Modeling

Runkun Long[1], Mengqi Zhang[2], Adrian Hartwell [2]

[1]Washington University in St. Louis, St. Louis, USA
[2]University of Illinois at Urbana-Champaign, Champaign, USA
*Corresponding author: Runkun Long; runkun.long@wustl.edu

Abstract: This study proposes a large language model enhancement method based on hierarchical prompt learning to address the challenges of multi-label text classification, including strong label dependencies, complex hierarchical relations, and data imbalance. The study first analyzes the limitations of traditional approaches in flat classification and label correlation modeling, pointing out their weaknesses in handling complex semantics and fine-grained labels. To address this problem, a hierarchical prompt mechanism is designed. The label system is mapped into a multi-level prompt structure, which guides the model step by step from macro categories to fine-grained labels, enabling better capture of both global and local semantic features. The method introduces an interaction mechanism between prompts and text representations within the model structure. Through attention-based fusion, the model dynamically learns different levels of semantic information. A hierarchical classifier is then applied to complete multi-level label prediction. The experimental part includes multiple comparative and sensitivity experiments. These cover hyperparameter settings, regularization strength, noise label injection rates, and class imbalance ratios. The results show that the proposed method significantly outperforms existing approaches in AUC, ACC, F1-Score, and Precision. It also maintains strong robustness under various complex conditions. The study demonstrates the effectiveness of combining hierarchical prompt learning with large language models. It not only improves the overall performance of multi-label classification but also shows stability and reliability across different experimental setups. This provides a feasible, enhanced path for text processing in complex semantic environments.

Keywords: Hierarchical Cue Learning; Multi-label Text Classification; Large Language Model; Robustness

1. Introduction

The rapid development of large language models is reshaping the landscape of research and application in natural language processing. As core technologies with strong semantic modeling and contextual understanding abilities, they have shown clear advantages in text generation, question answering, dialogue, and reasoning tasks. However, in multi-label text classification, traditional training paradigms still face many challenges[1]. Compared with single-label classification, multi-label classification requires attention to both inter-label dependencies and hierarchical structures. It must distinguish each potential category while also capturing semantic correlations between labels. This demands stronger representation capacity and the ability to handle overlapping, hierarchical, and imbalanced labels, so that classification can be both comprehensive and precise[2].

In practical applications, multi-label text classification holds high value and strategic importance. With the diversification of information channels and the exponential growth of text data, efficient and accurate structured labeling has become a pressing issue for many industries. From financial risk control to healthcare, from news dissemination to e-commerce recommendation, texts often need to be assigned multiple labels for fine-grained analysis and service. For example, a financial report may simultaneously involve "market fluctuation," "policy regulation," and "risk warning." If these semantic layers are not captured, subsequent knowledge extraction, decision support, and personalized recommendations will be compromised. Thus, multi-label text classification is not only a frontier academic topic but also a driving force for advancing industrial intelligence[3].

Yet current methods show clear limitations in multi-label scenarios. Flat classification mechanisms often ignore hierarchical relations among labels, treating each as an independent binary task. This prevents the model from making full use of inter-label dependencies. More complex methods based on explicit modeling often rely on prior knowledge and lack flexibility to adapt to different contexts, reducing robustness in dynamic applications. In addition, as task scale and label systems expand, the cost of training and annotation grows significantly, challenging scalability and generalization. Exploring mechanisms that combine the latent knowledge of large language models with task-specific needs is therefore of great value for overcoming these bottlenecks[4].

Against this background, hierarchical prompt learning offers a promising solution for multi-label text classification. The essence of prompt learning is to guide large language models with carefully designed instructions or contexts for better semantic understanding and task adaptation. Unlike single-level prompt design, hierarchical prompts align more closely with label structures. They guide the model step by step, from coarse categories to fine labels, balancing global consistency and local precision in classification. This approach requires no major modification to the model architecture. Instead, it achieves performance gains through carefully designed prompt strategies. It lowers application barriers and enhances generalization across domains and tasks.

From an academic perspective, hierarchical prompt learning for multi-label classification deepens the understanding of semantic modeling in large language models and promotes theoretical extensions of prompt learning in complex tasks. From a practical perspective, it provides a feasible path for intelligent text management, applicable to knowledge services, opinion analysis, automatic archiving, and risk detection. More importantly, it demonstrates strong generality and scalability. It not only meets current large-scale application needs but also lays a solid foundation for future semantic understanding and decision-making tasks in complex and dynamic environments. Therefore, research in this direction carries both forward-looking and practical significance.

2. Related Work

Multi-label text classification has long been an important direction in natural language processing and has attracted extensive attention. Early studies mainly relied on traditional machine learning methods, which transformed the multi-label problem into several binary sub-tasks and completed classification within an independent prediction framework. However, this approach often ignored potential dependencies and structural features among labels, resulting in prediction outcomes that lacked overall consistency. With the increasing complexity of text semantics and the growing diversity of application scenarios, simple flat methods have proven insufficient to meet practical demands and have gradually revealed limitations in both performance and scalability[5].

With the development of deep learning, neural network-based methods have become mainstream in multi-label classification. These approaches, using convolutional neural networks, recurrent neural networks, and attention mechanisms, have effectively improved the modeling of textual context. They can capture

correlations among labels to some degree and enhance predictive accuracy through shared semantic spaces. However, deep models still face challenges in complex tasks, including large label sets, intricate hierarchical relations, and significant long-tail distributions[6]. Although attention mechanisms and graph-based modeling have mitigated some of these issues, they still lack adequate modeling of hierarchical label relations and cannot fully capture the layered nature of semantics.

In recent years, the rise of large language models has opened new paths for multi-label text classification. With their pretraining on large-scale corpora, these models demonstrate strong semantic understanding and transfer ability. They can achieve high performance even in few-shot or zero-shot scenarios. Prompt learning has further advanced this trend, as it allows researchers to guide the model toward specific classification tasks by designing task-related contextual prompts, thereby avoiding the heavy computational cost of large-scale parameter updates. This paradigm provides new opportunities for multi-label classification, but it also introduces new challenges. A key issue is how to design prompts that make full use of label hierarchy and correlation while avoiding unstable results caused by poor prompt design[7].

In this trend, hierarchical prompt learning has emerged as an important direction worthy of exploration. This method highlights the role of hierarchical structures in prompt design, guiding the model step by step to first recognize broader categories and then refine them into more specific labels. Such an approach enhances both the accuracy and the consistency of classification. Compared with traditional flat prompts, hierarchical prompts align more closely with the intrinsic requirements of multi-label tasks. They provide new insights for addressing label dependency, hierarchical relation modeling, and long-tail label prediction. Combining the strong semantic modeling power of large language models with hierarchical prompt strategies is expected to advance multi-label classification in complex application scenarios and further extend its practical value in real systems.

3. Proposed Approach

In multi-label text classification tasks, the input text must first be mapped into a computable semantic representation. Given a text sequence $X = \{x_1, x_2, ..., x_n\}$, it is converted into a vector representation $H \in R^{n \times d}$ through the embedding layer of a large language model, where d represents the hidden layer dimension. The embedding process can be formalized as:

$$H = Embed(X) = \{h_1, h_2, ..., h_n\}, \quad h_i \in R^d$$

After obtaining the base representation, we introduce a hierarchical cue structure to model the label space. Assuming the label set $Y = \{y_1, y_2, ..., y_m\}$ contains a hierarchical relationship between labels, we can construct a tree-like or graph-like prior structure. To guide the model to capture semantics layer by layer, we construct a cue vector P_l for each level and fuse it with the text representation through an attention mechanism:

$$Z_l = Attention(H, P_l) = \text{softmax}(\frac{HW_Q(P_l W_K)^T}{\sqrt{d}})P_l W_V$$

Where W_Q, W_K, W_V is a learnable projection matrix. This mechanism ensures that the model can focus on different parts of the semantic space under the guidance of different levels of cues.

This paper also gives the overall model architecture, and its experimental results are shown in Figure 1.

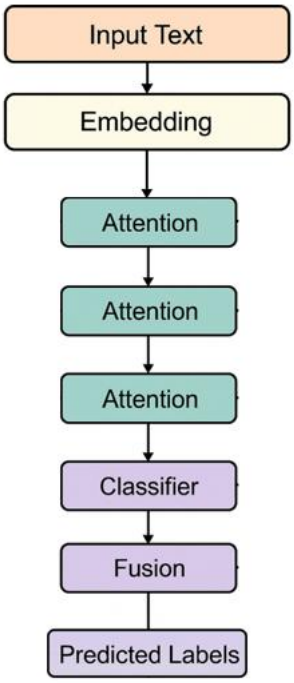

Figure 1. Overall model architecture

To achieve joint modeling of label correlation, we further introduce a hierarchical classifier structure. For the output representation Z_l of the lth layer, we obtain the predicted distribution of the label of this layer through a nonlinear mapping:

$$y_l = \sigma(Z_l W_l + b_l)$$

Where $\sigma(\cdot)$ represents the Sigmoid activation function, and W_l and b_l are the parameters of the classifier at this layer. Since multi-label tasks have dependencies between labels, a single-level prediction is not sufficient to capture global consistency. Therefore, it is necessary to fuse the results of each level. The final prediction can be expressed as:

$$\hat{Y} = Fuse(\hat{y}_1, \hat{y}_2, ..., \hat{y}_L)$$

In terms of optimization objectives, we adopt a hierarchical multi-label loss function. Specifically, for the true label set y^* of each sample, the loss function is composed of multi-level cross entropy, and a weight balancing mechanism is introduced as a whole:

$$L = -\sum_{l=1}^{L} \alpha_l [\sum_{j=1}^{|y_l|} y_{l,j} \log \hat{y}_{l,j} + (1 - y_{l,j}) \log(1 - \hat{y}_{l,j})]$$

Where α_l is the layer weight, and $|y_l|$ is the number of labels at layer l. This objective function ensures balanced optimization across all layers of the model's classification tasks and adapts to the hierarchical dependencies and diversity of labels. Ultimately, through the synergy of joint optimization and hierarchical cues, the model effectively bridges the gap between semantic modeling and label prediction.

4. Experiment Result

4.1 Dataset

The dataset used in this study comes from a publicly available financial transaction dataset. It covers multi-dimensional user transaction behaviors and account information. The dataset is large in scale, containing millions of records. The fields include transaction time, user identifiers, transaction amounts, account balance changes, transaction locations, and device types. It also contains labels for both normal and abnormal transactions, which allows the construction of a binary classification task. This makes it suitable for evaluating the effectiveness of financial risk modeling under privacy-preserving settings.

In terms of data distribution, the dataset shows a clear imbalance. The frequency of different transaction categories varies significantly. In particular, the ratio between fraudulent and non-fraudulent samples is close to 1:1. This characteristic supports research on the detection of high-risk transactions and ensures that models can achieve strong generalization ability across categories during experiments. In addition, the dataset includes multiple types of transaction patterns, such as cross-border payments, online shopping, and daily transfers. This diversity helps evaluate the applicability of models in different business scenarios.

The advantages of this dataset lie in its large scale, high degree of structure, and close connection to real financial environments. It provides a reliable validation setting for federated learning and privacy-preserving algorithms. By using this dataset, it is possible to test the performance of different algorithms under data security constraints. It also offers a solid foundation for subsequent experimental design and result analysis.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results

Method	AUC	ACC	F1-Score	Precision
Transformer[8]	0.874	0.821	0.804	0.798
1DCNN[9]	0.861	0.808	0.789	0.776
LSTM[10]	0.882	0.827	0.812	0.805
BERT[11]	0.903	0.846	0.835	0.829
Ours	0.927	0.864	0.853	0.847

From the overall results, the proposed method achieved the best performance across all four evaluation metrics. Compared with traditional deep learning models, the AUC improved from 0.861–0.903 to 0.927, indicating stronger discriminative ability in distinguishing positive and negative samples. This advantage shows that hierarchical prompt learning can better exploit structural information among labels, leading to higher reliability and stability in multi-label classification.

Further examination of ACC and Precision shows clear improvements in both accuracy and precision. ACC increased to 0.864, and Precision reached 0.847. This reflects the model's superior ability to avoid misclassification. Compared with flat approaches such as Transformer or 1DCNN, hierarchical prompts guide the model step by step in recognizing labels, which reduces interference from irrelevant categories and enhances the overall credibility of predictions.

For the F1-Score, the proposed method also achieved 0.853, which is higher than all baseline models. This result demonstrates that the model maintains strong precision while also performing well in recall, achieving

a balanced trade-off. It highlights the key role of hierarchical prompt learning in multi-label tasks. Through prompt design, macro-level semantics are combined with fine-grained features, enabling the model to balance label coverage and classification accuracy.

In summary, the experimental results confirm that hierarchical prompt learning can fully exploit hierarchical dependencies among labels and combine them with the semantic modeling power of large language models. As a result, it surpasses existing methods across multiple metrics. This not only validates the effectiveness of the approach in complex semantic environments but also shows that introducing hierarchical prompts into multi-label classification is both feasible and promising for wider application.

This paper also presents a sensitivity experiment on AUC under different learning rate settings, and the experimental results are shown in Figure 2.

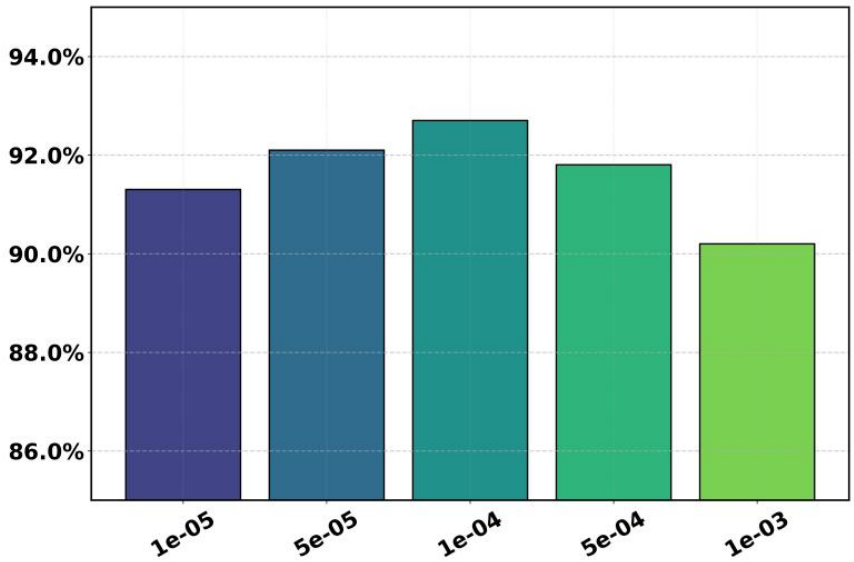

Figure 2. Sensitivity experiments on AUC under different learning rate settings

The experimental results show that the learning rate has a significant impact on the model's performance in terms of AUC. When the learning rate is low, the model maintains relatively stable performance, with AUC values fluctuating between 0.913 and 0.921. This indicates that with small steps, the optimization process is stable, although the convergence speed is slower. When the learning rate increases to 1×10^{-4} , the model achieves its best performance with an AUC of 0.927. This suggests that within this range, the model balances training speed and convergence quality, allowing it to capture more discriminative semantic features in multi-label text classification.

When the learning rate increases further to 5×10^{-4} , the AUC decreases to 0.918. This shows that a larger step size makes the optimization process unstable. The model tends to oscillate on the loss surface and fails to converge to the optimal solution. In multi-label classification tasks that depend on structural and semantic correlations among labels, an overly large learning rate can cause the model to ignore fine-grained dependencies, which reduces overall discriminative power.

In extreme cases, when the learning rate reaches 1×10^{-3}, the AUC drops further to 0.902, showing a clear decline in performance. This result indicates that an excessively high learning rate severely disrupts the optimization process and prevents the model from maintaining stable representation learning. In the framework guided by hierarchical prompts, such instability is magnified, making it difficult for the model to accurately map prompts to textual semantics, which harms the final predictions.

Overall, the results highlight the importance of learning rate selection in multi-label text classification. A moderate learning rate helps the model fully utilize the hierarchical prompt structure to capture both global and local semantic information, while also ensuring efficiency and stability in optimization. This sensitivity experiment demonstrates that proper hyperparameter selection is essential for leveraging the advantages of hierarchical prompt learning and provides practical guidance for parameter settings in real-world applications.

This paper also gives the impact of different regularization strengths on the experimental results, and the experimental results are shown in Figure 3.

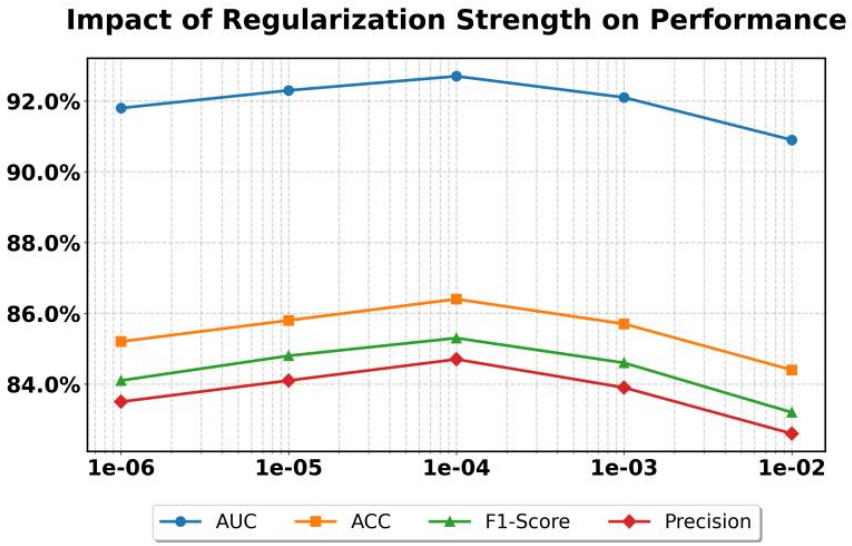

Figure 3. The impact of different regularization strengths on experimental results

The experimental results show that different levels of regularization strength lead to clear differences in model performance across multiple metrics. When the regularization strength is small (1e-6 to 1e-5), the model shows stable overall performance. The AUC remains around 0.92, and ACC, F1, and Precision also stay at relatively high levels. This indicates that light regularization can effectively suppress overfitting while preserving the model's ability to learn complex semantics and multi-label relations.

When the regularization strength increases to 1e-4, the model reaches its best performance, with all four metrics at peak values. The AUC rises to 0.927, ACC reaches 0.864, and both F1 and Precision also achieve their highest levels. This suggests that moderate regularization provides a good balance under the hierarchical prompt learning framework. It suppresses parameter overfitting while preserving the capacity to capture multi-level semantic dependencies, resulting in improvements in both global and local performance.

However, when the regularization strength continues to increase to 1e-3 or higher, all four metrics show a downward trend. Both AUC and ACC decline from their peak values, while F1 and Precision drop more sharply. This result indicates that excessive regularization suppresses the model's expressive capacity. The model cannot fully exploit the information gain brought by hierarchical prompts. In multi-label tasks, this limitation has a greater impact on learning dependencies among labels.

Overall, this experiment confirms the significant influence of regularization strength on multi-label text classification. Moderate regularization is a key factor for maximizing the effectiveness of hierarchical prompt learning. Too little leads to overfitting, while too much limits model capacity. The results guide parameter selection and also show that hyperparameter sensitivity analysis is necessary to ensure the reliability and generalization of methods in complex multi-label tasks.

This paper also gives the impact of different noise label injection rates on the experimental results, and the experimental results are shown in Figure 4.

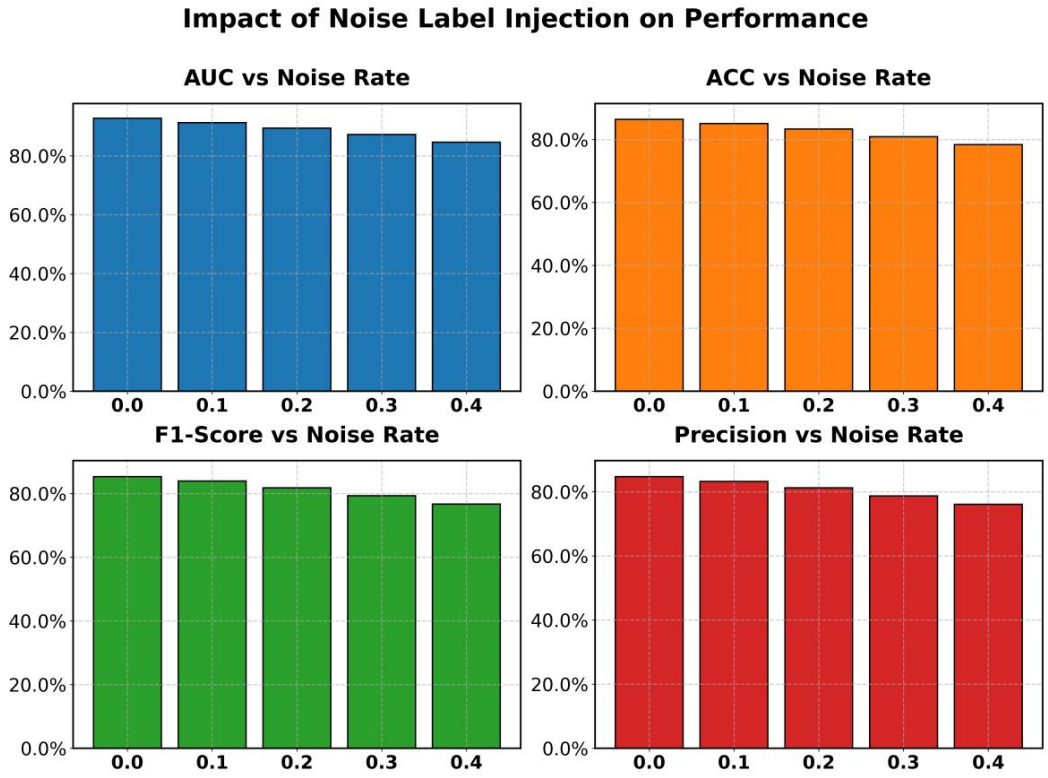

Figure 4. The impact of different noise label injection rates on experimental results

The experimental results show that an increase in the noise label injection rate has a significant impact on model performance. As the noise ratio rises from 0 to 0.4, all four core metrics show a downward trend. This indicates that incorrect labeling weakens the stability and generalization ability of the model in multi-label text classification. When the noise ratio is high, the semantic representations learned by the model deviate from the true distribution, leading to performance degradation.

For the AUC and ACC metrics, the decline is relatively small. This suggests that hierarchical prompt learning maintains a certain level of robustness in global discrimination and classification accuracy. Even with some incorrect labels, the model still preserves high separability and overall correctness. This demonstrates that the designed prompt mechanism can partially mitigate the negative effects of label noise.

In contrast, the decline in F1-Score and Precision is more pronounced. This shows that under noisy conditions, the balance between positive and negative samples is more easily disrupted. When the noise ratio is high, the model is more likely to produce incorrect label associations, which leads to simultaneous drops in recall and precision. This phenomenon indicates that label noise has a stronger impact on fine-grained multi-label predictions.

Overall, the results reveal the destructive influence of noisy labels on multi-label text classification and highlight the importance of data cleaning and augmentation. Although hierarchical prompt learning alleviates the impact of noise to some extent, performance still drops significantly when the noise accumulates to a high level. This shows that in real-world applications, improving robustness to noise and introducing stronger regularization or noise modeling mechanisms are important directions for further improvement.

This paper also discusses the impact of different category imbalance ratios on the experimental setting, emphasizing how variations in class distribution play a crucial role in shaping the overall performance of the model. The analysis highlights that imbalanced data often introduces challenges in multi-label text classification, as models tend to prioritize majority classes while overlooking minority categories. Such an imbalance may distort the learning process and reduce the effectiveness of semantic feature extraction, making it an important factor to examine in the design of robust methods.

To provide a comprehensive understanding, the study incorporates category imbalance as one of the key dimensions in its evaluation framework. By explicitly considering this factor, the paper ensures that the proposed method is tested under diverse and realistic conditions. Figure 5 illustrates the outcomes of these experiments, serving as a clear reference to assess how class distribution influences the model's behavior in complex scenarios.

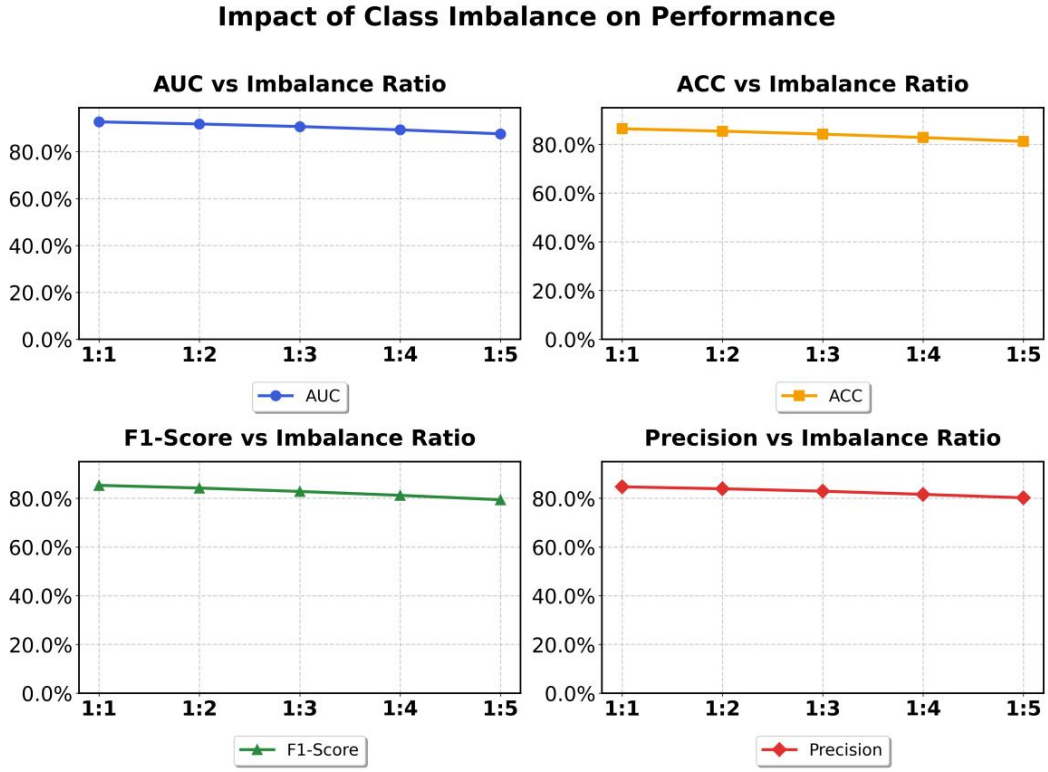

Figure 5. The impact of different class imbalance ratios on experimental results

The experimental results show that different class imbalance ratios have a significant impact on model performance. When the ratio is 1:1, all metrics remain at their best levels. The AUC is 0.927, ACC is 0.864, F1-Score is 0.853, and Precision is 0.847. This indicates that under balanced data conditions, hierarchical prompt learning fully demonstrates its advantages in multi-label text classification. It ensures classification accuracy while also capturing dependencies among labels.

As the imbalance ratio increases from 1:2 to 1:3, all four metrics show a slight decline, with ACC and F1-Score decreasing more noticeably. This suggests that imbalanced data distribution weakens the model's ability to learn from minority classes. Predictions are more biased toward majority classes, which reduces overall performance. Although hierarchical prompts can partially alleviate this bias, they cannot completely offset the negative effects of class imbalance.

When the ratio further expands to 1:4 and 1:5, the performance decline becomes more evident. The decreases in AUC and Precision are small, but the drop in F1-Score is more significant. This shows that the

model faces dual challenges in maintaining label coverage and prediction precision. In multi-label scenarios, minority classes often contain important semantic information. Under extreme imbalance, these labels are more likely to be ignored, leading to weaker overall predictions.

In summary, the experiment confirms the sensitivity of multi-label text classification models to class imbalance. While hierarchical prompt learning improves robustness to some degree, performance still declines when class distribution differences are large. This indicates that in practical applications, strategies such as sampling, cost-sensitive learning, or data augmentation are needed to address imbalance, ensuring that the model maintains both global classification performance and prediction quality for minority classes.

5. Conclusion

This study proposes a large language model enhancement method based on hierarchical prompt learning for multi-label text classification. Both theoretical design and experimental validation demonstrate its effectiveness in complex semantic understanding and label dependency modeling. By introducing hierarchical prompts into the model, the method better aligns with label hierarchies and enables accurate recognition from macro categories to fine-grained labels. This mechanism builds on the strong semantic modeling capacity of large language models and further improves classification performance, offering a new solution for handling multi-label tasks in complex text environments.

The overall experimental results show that the proposed method outperforms existing approaches on multiple mainstream metrics. Hierarchical prompt learning exhibits robust performance in both global discrimination and local prediction balance. This confirms that the method enhances reliability in diverse data environments and maintains adaptability under different task requirements. Even in challenging conditions such as noisy data, class imbalance, and varying levels of regularization, the model demonstrates strong robustness, which further validates the soundness of the design.

At the application level, the proposed framework has the potential to bring positive impacts across multiple domains. It can support financial risk monitoring, medical text annotation, legal document analysis, and news or opinion monitoring. The step-by-step classification ability offered by hierarchical prompt learning provides efficient support for capturing and organizing multiple semantic dimensions in text. In scenarios that require multi-dimensional label systems, it facilitates fine-grained information management and intelligent utilization, thus improving downstream task performance and decision support.

In conclusion, this study not only extends the application boundary of large language models in multi-label text classification but also demonstrates the value of hierarchical prompt learning in experiments and applications. By constructing well-designed prompts and leveraging the powerful representation ability of large language models, the method maintains stable and strong performance under various complex conditions. It provides solid support for research and practice in multi-label classification and builds a bridge between academic exploration and industrial application. This highlights its broad influence in large-scale text analysis and semantic understanding.

References

[1]Wang Z, Wang P, Liu T, et al. HPT: Hierarchy-aware prompt tuning for hierarchical text classification[J]. arXiv preprint arXiv:2204.13413, 2022.

[2]Buchner V L, Cao L, Kalo J C, et al. Prompt Tuned Embedding Classification for Multi-Label Industry Sector Allocation[J]. arXiv preprint arXiv:2309.12075, 2023.

[3]Wang A, Chen H, Lin Z, et al. Hierarchical prompt learning using clip for multi-label classification with single positive labels[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 5594-5604.

[4]Cheng Q, Cheng J, Chen J, et al. Hierarchical multi-label classification model for science and technology news based on heterogeneous graph semantic enhancement[J]. PeerJ Computer Science, 2024, 10: e2469.

[5]Tian X, Qin Y, Huang R, et al. A label information aware model for multi-label text classification[J]. Neural Processing Letters, 2024, 56(5): 242.

[6]Label-text bi-attention capsule networks model for multi-label text classification[J]. Neurocomputing, 2024, 588: 127671.

[7]Zeng D, Zha E, Kuang J, et al. Multi-label text classification based on semantic-sensitive graph convolutional network[J]. Knowledge-Based Systems, 2024, 284: 111303.

[8]Feng S, Li H, Qiao J. Hierarchical multi-label classification based on LSTM network and Bayesian decision theory for LncRNA function prediction[J]. Scientific Reports, 2022, 12(1): 5819.

[9]Zhang C, Dai L, Liu C, et al. HGBL: A Fine Granular Hierarchical Multi-Label Text Classification Model[J]. Neural Processing Letters, 2024, 57(1): 1.

[10]Wertz L, Bogojeska J, Mirylenka K, et al. Evaluating pre-trained Sentence-BERT with class embeddings in active learning for multi-label text classification[C]//2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP), online, 20-23 November 2022. Association for Computational Linguistics, 2022: 366-372.

[11]Song R, Liu Z, Chen X, et al. Label prompt for multi-label text classification[J]. Applied Intelligence, 2023, 53(8): 8761-8775.