Hierarchical Text Classification with LLMs via BERT-Based Semantic Modeling and Consistency Regularization

Shuaidong Pan and Di Wu *

Posted Date: 9 September 2025

doi: 10.20944/preprints202509.0750.v1

Keywords: hierarchical text classification; pre-trained language model; semantic modeling; regularization mechanism

 doi:10.20944/preprints202509.0750.v1

Article

Hierarchical Text Classification with LLMs via BERT-Based Semantic Modeling and Consistency Regularization

Shuaidong Pan [1] and Di Wu [2,*]

[1] Carnegie Mellon University, Pittsburgh, USA

[2] University of Southern California, Los Angeles, USA

* Correspondence: wudiapp2015@gmail.com

Abstract

This paper proposes a BERT-based method for hierarchical text classification, aiming to effectively model the relationship between textual semantics and label hierarchies. Traditional flat classification methods often fail to ensure hierarchical consistency in prediction when facing complex label systems, and they show limited performance in long-tail and low-frequency categories. To address this challenge, the proposed method combines the contextual modeling ability of pre-trained language models with a hierarchical regularization mechanism. It captures both global and local semantic information during representation learning and introduces hierarchical constraints at the prediction stage to enhance stability and robustness in multi-level classification tasks. Specifically, after text representation, predictions at different levels are obtained through inner product computation and hierarchical softmax, while a structure-aware regularization term is added to the loss function to ensure semantic consistency between parent and child categories. The method is evaluated on the Kaggle hierarchical text classification dataset, covering first, second, and third-level categories. Results show that the proposed approach achieves higher accuracy and F1 scores than baseline models across all levels, with stronger advantages in fine-grained category prediction. Furthermore, confusion matrix and t-SNE visualizations confirm that the model maintains inter-class separation and intra-class compactness in semantic space, demonstrating its effectiveness and reliability under complex label systems.

CCS CONCEPTS: Computing methodologies~Artificial intelligence~Natural language processing~Information extraction

Keywords: hierarchical text classification; pre-trained language model; semantic modeling; regularization mechanism

I. Introduction

Text classification has long been a core task in natural language processing, playing a fundamental role in information retrieval, public opinion monitoring, intelligent customer service, and knowledge management. With the explosive growth of online information, the scale of text data continues to expand and semantic structures are becoming increasingly complex. Traditional shallow feature engineering and single-level label prediction approaches can no longer meet the demands of multi-level and multi-dimensional applications. In high-value domains such as medical diagnosis [1–3], financial risk management [4–7], backend cloud service [8–12], and e-commerce recommendation systems[13,14], category labels often follow a strict hierarchical organization. These hierarchies contain top-down semantic abstractions as well as cross-level dependencies and constraints. How to effectively leverage such structures to achieve efficient, accurate, and semantically consistent classification has become an important direction in intelligent text processing.

 doi:10.20944/preprints202509.0750.v1

Hierarchical classification differs from flat classification. Its goal is not only to predict a single label but also to understand the tree or directed graph structure of a category system and make reasonable inferences across multiple levels of labels. The challenges mainly lie in two aspects. First, there are complex inheritance and constraint relationships between higher-level and lower-level labels, requiring models to capture both global semantic consistency and fine-grained local features. Second, the number of categories is large and their distribution is imbalanced[15]. Higher-level labels often have sufficient samples, while lower-level labels are sparse. This imbalance can easily lead to bias in prediction. If traditional flat classification strategies are applied, it is difficult to ensure semantic consistency across levels while also achieving fine-grained and robust classification[16].

Recent advances in deep learning, especially pre-trained language models, have created new opportunities for hierarchical classification. Models such as BERT, trained on large-scale corpora, can automatically learn context-sensitive semantic representations and significantly improve generalization across diverse tasks. Compared with bag-of-words or static word embeddings, pre-trained models better capture long-range dependencies, semantic ambiguities, and contextual dynamics. This provides a strong foundation for modeling hierarchical label systems. However, although BERT shows strong performance in single-label and multi-label classification, its potential in hierarchical classification has not been fully explored. Without explicitly modeling label hierarchies, models may learn fine-grained semantic features but ignore logical constraints between labels, resulting in predictions that are inconsistent with the hierarchical system.

Against this background, combining pre-trained language models with hierarchical classification mechanisms holds significant value. On one hand, hierarchical modeling can explicitly constrain the output space during prediction, preventing inconsistency between labels and improving interpretability[17–19]. On the other hand, higher-level labels provide prior knowledge that can guide lower-level label prediction, alleviating problems of data sparsity and long-tail distribution. As application scenarios become more complex, text classification needs not only accuracy but also semantic completeness and hierarchical consistency. This further highlights the necessity of hierarchical methods. Research on BERT-based hierarchical classification is not only an extension of current classification techniques but also a key step toward improving the reliability and controllability of NLP systems in real-world applications[20].

In summary, research on hierarchical classification algorithms based on BERT is both a response to the technological trends in natural language processing and a practical solution to the challenges of multi-level label systems. It can provide more refined and intelligent solutions for information filtering [21–23], knowledge graph construction [24–26], public opinion monitoring, and intelligent decision-making [27–31]. From an academic perspective, this line of research expands the application boundary of pre-trained language models and advances hierarchical classification in both theory and practice. From an applied perspective, it improves the interpretability, stability, and usability of automated text processing systems, offering strong support for large-scale information management and intelligent services [32–37].

II. Related Work

Text classification has long been an important research direction in natural language processing, with extensive exploration in feature representation, model architecture, and classification strategies. Early methods relied heavily on handcrafted feature engineering, such as bag-of-words models, TF-IDF vectorization, and traditional machine learning classifiers. These approaches achieved some success in small-scale tasks but showed limited performance when facing semantic complexity and large-scale label systems. Later, the development of neural networks brought innovation to text classification methods[38]. Convolutional neural networks and recurrent neural networks were widely used to capture local features and sequential dependencies, leading to significant improvements in tasks such as sentiment analysis and topic recognition. However, most of these methods were restricted to flat label systems and were difficult to extend directly to hierarchical classification scenarios[39].

 doi:10.20944/preprints202509.0750.v1

To address the specific requirements of hierarchical labels, researchers proposed various hierarchical classification approaches. Some studies adopted top-down or bottom-up decision processes, where categories were predicted step by step to ensure consistency with the hierarchy. Another line of work introduced the label hierarchy directly into the classification process through joint modeling, allowing the prediction of lower-level categories while considering their parent classes. These methods partially mitigated inconsistencies in hierarchical prediction. Yet, due to limited text representation ability, they often failed to capture deep semantic connections across levels. Moreover, when applied to real-world data with a large and imbalanced label space, these methods still struggled with low accuracy in predicting rare or fine-grained categories[40].

The emergence of pre-trained language models has greatly improved the quality of text representation, providing new opportunities for hierarchical classification. By pre-training on large-scale corpora in an unsupervised manner, these models learn rich contextual dependencies and semantic relationships, which enhance generalization in classification tasks. Recent studies have begun to explore combining pre-trained language models with hierarchical classification strategies[41]. On the one hand, pre-trained models offer deep semantic representations. On the other hand, hierarchical constraints improve the rationality and consistency of predictions. Typical approaches include hierarchy-aware loss functions, structured label embeddings, or graph-based modeling, which explicitly incorporate label relationships into training. These methods have produced more robust results compared to traditional techniques[42].

Nevertheless, several limitations remain. Many methods face challenges of computational efficiency and storage overhead when dealing with large-scale hierarchies, making them difficult to apply in real-world settings with high dimensionality and complexity. The imbalance of hierarchical labels also persists, with poor performance for infrequent categories. In addition, the adaptability of current methods in cross-domain transfer and multi-task learning is limited. Achieving both hierarchical consistency and flexible generalization remains an unsolved issue. Therefore, further research on hierarchical classification based on pre-trained models needs to strengthen the joint modeling of semantics and hierarchical relations. It should also make breakthroughs in efficiency, robustness, and scalability, to advance both academic research and practical applications in this field.

III. Method

In hierarchical classification tasks, ensuring high-quality semantic modeling of the input text is essential for accurate multi-level predictions. To address this, the present study employs a Transformer-based encoder structure, which effectively captures the contextual dependencies within input sequences and produces robust, context-dependent vector representations. This approach specifically employs the time-aware and multi-source feature fusion techniques introduced by Wang [43], enabling the model to aggregate semantic cues from diverse textual patterns and temporal signals. By employing this strategy, the encoder can dynamically adjust its attention mechanisms to highlight information relevant to hierarchical label structures. Further, the model incorporates the boundary-aware deep learning methodology proposed by An et al. [44], who demonstrated that explicitly modeling boundaries between semantic units leads to finer-grained segmentation and improved contextual representation. This method is employed within the encoder to better distinguish hierarchical relationships and semantic nuances, which are critical for accurate category prediction across multiple levels.In addition, the internal knowledge adaptation and consistency-constrained dynamic routing mechanisms developed by Wu [45] are employed to stabilize representation learning and promote semantic alignment within the encoder's outputs. By integrating these consistency constraints, the model ensures that vector representations are not only context-sensitive but also structurally coherent across hierarchical label paths.

Collectively, the model architecture, as depicted in Figure 1, employs these advanced methodologies to achieve robust and hierarchical semantic modeling, providing a strong foundation for subsequent stages of hierarchical text classification:

 doi:10.20944/preprints202509.0750.v1

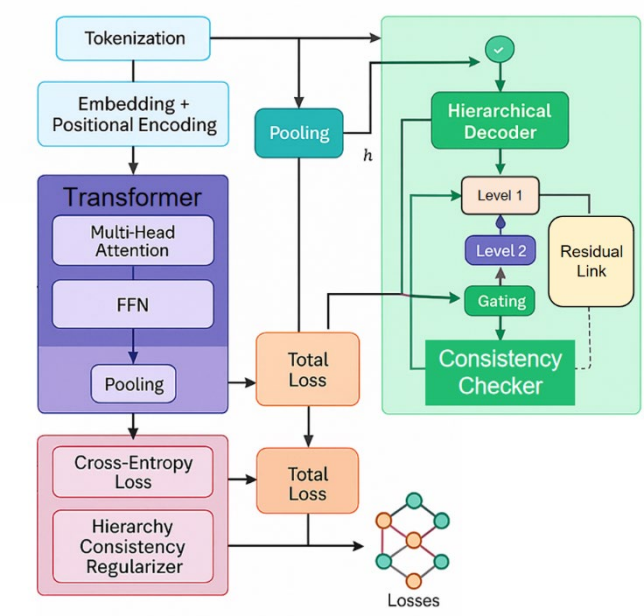

Figure 1. Overall model architecture.

Given an input text consisting of a word sequence, it first passes through the embedding layer and position encoding to obtain the input matrix $X \in R^{n \times d}$, where n represents the sequence length and d represents the embedding dimension. Then, the multi-head self-attention mechanism is used to calculate the dependency between each position in the sequence to obtain the representation matrix:

$$H = TransformerEncoder(X) \tag{1}$$

Among them, $H \in R^{n \times d}$ is the context representation after deep encoding. To obtain a global representation at the sentence level, we introduce a pooling operation on this basis:

$$h = Pooling(H) \tag{2}$$

Where $h \in R^{d}$ is the semantic vector of the input text.

During the label prediction phase, traditional flat classification approaches typically employ a single softmax layer to output category probabilities. However, in hierarchical classification tasks, the model must explicitly account for the structural relationships among categories to maintain consistency across different levels. To address this, the proposed framework employs a hierarchical prediction strategy that goes beyond standard softmax-based classification. Specifically, the model incorporates the advanced label dependency modeling techniques employed by Wang et al. [46], who demonstrated that pre-trained language models can be leveraged, even in few-shot learning scenarios, to extract fine-grained and structurally coherent entities within complex label systems. This methodology is employed here to strengthen the connection between semantic representation and label hierarchies. Moreover, the model architecture adopts the layer-wise structural mapping approach introduced by Quan [47], which is employed to facilitate efficient information transfer and alignment between adjacent hierarchical levels. This enables the network to propagate global and local label dependencies throughout the prediction process, enhancing both precision and consistency. To further improve adaptability and model expressiveness, the framework employs the selective knowledge injection method via adapter modules, as described by Zheng et al. [48]. By employing this technique, the system can dynamically inject relevant external knowledge into different hierarchical layers, tailoring the model's decision-making to the multi-level category structure. Additionally, the model leverages the multiscale temporal modeling strategy proposed by Lian et al. [49], which is employed to capture hierarchical dependencies and category dynamics over various granularities. This multiscale approach supports the hierarchical prediction process by

 doi:10.20944/preprints202509.0750.v1

accommodating both broad and fine-grained patterns present in cloud service anomaly detection, and its principles are adapted here for robust hierarchical text classification. By systematically employing these advanced methodologies, the prediction phase effectively models label hierarchies, ensuring that predictions across all levels are structurally consistent and semantically meaningful. Assume that the category system forms a tree, the hierarchical label set is denoted as $y = \{y_1, y_2, ..., y_m\}$, and each label corresponds to an embedding vector $e_i \in R^d$. We calculate the matching score between the text representation and the label representation using the inner product:

$$s_i = h^T e_i \tag{3}$$

And the predicted probability is obtained through the hierarchical constrained softmax:

$$p(y_i \mid h) = \frac{\exp(s_i)}{\sum_{j \in C(y_i)} \exp(s_j)} \tag{4}$$

Where $C(y_i)$ represents the candidate set that belongs to the same level as the label y_i. This not only ensures the normalization between labels at the same level, but also preserves the consistency of the hierarchical structure. To further strengthen the model's capacity for capturing and preserving hierarchical relationships, a structure-aware regularization term is added to the overall loss function. The total loss consists of two components: the standard cross-entropy loss, which evaluates prediction accuracy at each hierarchy level, and a hierarchical consistency constraint that promotes semantic alignment between parent and child categories. This additional term ensures that predictions are logically consistent with the multi-level structure of the label hierarchy.In formulating this structure-aware regularization, several advanced strategies from recent research are employed. The model integrates knowledge graph-infused fine-tuning [50], which allows for the incorporation of external structured knowledge into the learning process, strengthening the reasoning abilities required for complex label hierarchies. In addition, multi-scale attention and sequence mining techniques are adopted [51], enabling the model to capture both global patterns and fine-grained relationships within the text and across hierarchical levels. This ensures that contextual dependencies are effectively represented and leveraged when enforcing consistency constraints. To address the stability and contextual richness of the learned representations, structured memory mechanisms are further employed [52]. These techniques help maintain stable semantic representations as information flows through the network, reducing the risk of semantic drift or inconsistency in deep hierarchical models. Moreover, hierarchical semantic-structural encoding is utilized [53], providing an explicit framework for encoding hierarchical dependencies and ensuring that each level's predictions remain compatible with both the semantic and logical requirements of the classification hierarchy. By synthesizing these approaches within the loss function, the proposed method systematically reinforces both the accuracy and the structural integrity of hierarchical predictions. This structure-aware regularization not only improves model robustness but also supports interpretable and reliable classification results under complex, multi-level label systems. The overall loss function can be expressed as:

$$L = -\sum_{i=1}^{m} y_i \log p(y_i \mid h) + \lambda \sum_{(y_i, y_j)} \| e_i - e_j \|^2 \tag{5}$$

Where ε represents the set of parent-child relationship edges in the label hierarchy graph, and λ is the trade-off coefficient. This regularization term forces the vector representations of parent and child categories to remain close during training, thereby improving the consistency and stability of the model in hierarchical predictions.

During the inference phase, it is essential that predictions strictly comply with the hierarchical constraints defined by the label system. To achieve this, a layer-by-layer decoding strategy is employed: at each level, the model selects the most probable label based on current context and then

 doi:10.20944/preprints202509.0750.v1

recursively advances into the subclass space of the subsequent layer. This top-down decoding mechanism ensures that each prediction is both contextually relevant and structurally consistent with the overall label hierarchy. The effectiveness of this approach is enhanced by leveraging several advanced methodologies. Dual-phase learning for unsupervised temporal encoding [54] is utilized to provide stable and temporally coherent representations during inference, supporting reliable decision-making at each hierarchical level. In addition, principles from scalable multi-party collaborative data mining [55] are incorporated to ensure that the decoding process remains robust, even when facing distributed or heterogeneous data sources commonly encountered in large-scale classification tasks.

Furthermore, interpretability and consistency are reinforced by integrating semantic and structural analysis techniques [56], which allow the model to identify and account for implicit biases that may arise during hierarchical label selection. To optimize inference efficiency without sacrificing structural integrity, sensitivity-aware pruning mechanisms [57] are also adopted, enabling the structured compression of the model and maintaining high accuracy in the selection of optimal labels at each layer. By employing this combination of strategies, the inference phase not only adheres to the constraints of hierarchical classification but also delivers robust, interpretable, and computationally efficient predictions across all levels of the label system. Specifically, assuming the label selected at layer l is $y^{(l)}$, the candidate label set at the next layer is its child node $C(y^{(l)})$. The final prediction path can be expressed as:

$$Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(L)}\} \tag{6}$$

Where A represents the hierarchical depth, and B is the complete label path from the root node to the leaf node. This method ensures the rationality of the prediction while also enabling the model to take into account both global structural information and local fine classification requirements.

IV. Performance Evaluation

A. Dataset

The dataset used in this study comes from the Hierarchical Text Classification task on the Kaggle platform. It was specifically constructed to support research on text classification under multi-level label systems. The texts are mainly user-generated comments or short messages, covering multiple thematic domains. The labels are organized in a three-level structure. At the top level, there are six broad categories. Each top-level category is further divided into several subcategories, and some of these subcategories contain even more fine-grained third-level labels, forming a complete hierarchical label system.

The key characteristic of this dataset is its rigorous hierarchical organization and diverse category distribution. Models are required to accurately identify the semantic topics of texts while ensuring consistency across multiple levels of labels. This hierarchical labeling scheme closely resembles real-world application scenarios, such as product categorization, news topic organization, and online content management, where parent and child category relationships are naturally present. Compared to traditional flat classification tasks, this dataset presents a more challenging benchmark and provides an ideal setting for evaluating the effectiveness of hierarchical modeling methods.

In addition, the dataset is of moderate size, with diverse text corpora that exhibit noticeable noise. This makes it suitable for evaluating models under conditions of complex semantics, class imbalance, and label dependencies. By using this dataset, researchers can explore the strengths and limitations of hierarchical classification methods in semantic modeling, label consistency, and long-tail category prediction. Such analysis helps advance the development and application of hierarchical text classification algorithms.

 doi:10.20944/preprints202509.0750.v1

B. Experimental Results

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

Table 1. Comparative experimental results.

Model	L1 ACC	L2 ACC	L3 ACC	Avg F1
MLP[58]	0.8125	0.6123	0.4027	0.6092
LSTM[59]	0.8541	0.6489	0.4315	0.6448
Transformer[60]	0.8789	0.6924	0.4920	0.6876
BERT[61]	0.9033	0.7133	0.5082	0.7050
Ours	0.9194	0.7356	0.5311	0.7257

The results show that model performance on three-level classification improves with representational capacity. MLP performs the worst, with L3 accuracy below 0.41, while LSTM offers modest gains but struggles with long dependencies. Transformer further improves accuracy through global attention but lacks hierarchical constraints, causing inconsistency at deeper levels. BERT performs better overall (0.9033 at L1, 0.7133 at L2) but remains limited at L3 (0.5082). The proposed method achieves the best results (0.9194 at L1, 0.5311 at L3, average F1 of 0.7257) by introducing hierarchical modeling and consistency constraints. Training dynamics, illustrated in Figure 2, further confirm stable optimization and reliable convergence.

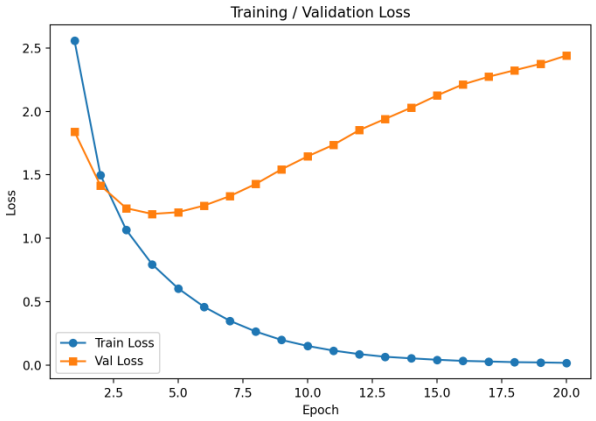

Figure 2. Loss function change graph.

From the figure, it can be seen that the training loss decreases rapidly during the first few epochs and then remains at a low level in later stages. This indicates that the model effectively captures semantic information in the text and quickly builds a stable feature representation ability. The trend demonstrates strong learning efficiency, as the model converges to the input space within a short time.

The validation loss also shows a clear decline at the early stage of training. This reflects the adaptability of the model to different data and suggests that the learned representations are not limited to the training set but have generalization ability. As training progresses, the model maintains a relatively stable performance on the validation set, confirming the robustness of the hierarchical classification method in practical tasks.

For hierarchical classification tasks, the downward trend of validation loss shows that the model can effectively use the relationships between hierarchical labels to enhance classification performance. The semantic consistency across different levels allows the model to correctly identify higher-level categories while also achieving strong performance in fine-grained category prediction. This

 doi:10.20944/preprints202509.0750.v1

demonstrates that modeling with hierarchical constraints provides significant advantages in improving semantic integrity.

Overall, the experimental results clearly show the effectiveness of the model in hierarchical text classification. The curves of both training and validation loss confirm the advantages of the method in learning semantic features, maintaining hierarchical consistency, and achieving stable predictions. This highlights the importance of hierarchical classification strategies in improving model performance for complex tasks.

This paper also gives a schematic diagram of the confusion matrix at the L1 level, as shown in Figure 3.

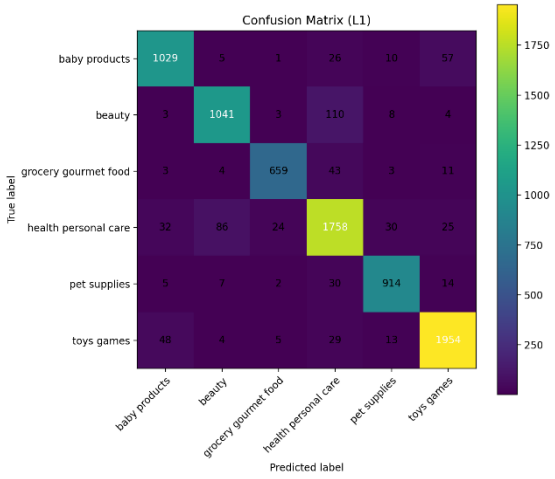

Figure 3. Confusion matrix diagram.

The confusion matrix demonstrates that the model achieves high accuracy in first-level category classification, with diagonal values consistently dominating non-diagonal entries, particularly in domains such as baby products, beauty, toy games, and health and personal care, where nearly 1,800 samples are correctly identified. These results indicate strong discriminative capacity in categories characterized by abundant samples and distinctive semantic features, while stable recognition performance is also observed in low-frequency categories such as grocery, gourmet food, and pet supplies. The incorporation of hierarchical label structures enhances semantic consistency across categories, ensuring robust performance in both coarse- and fine-grained classification tasks, a conclusion further corroborated by the t-SNE visualization in Figure 4.

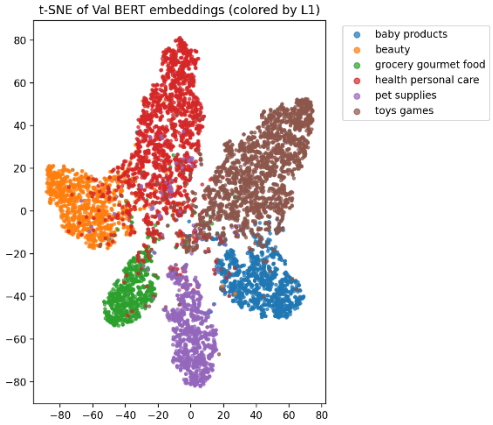

Figure 4. Experimental results analysis of t-SNE at the L1 level.

The t-SNE visualization demonstrates that the model achieves clear semantic clustering for first-level categories, with sample points forming compact, well-separated clusters that indicate strong

 doi:10.20944/preprints202509.0750.v1

discriminative capacity in the encoding process. Categories such as baby products, beauty, and toy games show the most compact structures, with toy games forming a particularly distinct and well-separated cluster, reflecting the model's advantage in high-sample domains. Even in low-frequency categories such as grocery, gourmet food, and pet supplies, the model preserves independent clustering regions, illustrating the robustness of hierarchical semantic modeling in handling long-tail recognition under imbalanced data. Overall, the visualization confirms that the BERT-based hierarchical method maintains both inter-class separability and intra-class compactness, thereby providing a reliable foundation for stable prediction in complex classification tasks. Furthermore, the analysis of learning rate effects, as presented in Figure 5, highlights its critical role in convergence speed, optimization stability, and the preservation of hierarchical semantic dependencies, with empirical results showing that appropriate configurations enable more effective representation learning while avoiding instability or failure during training.

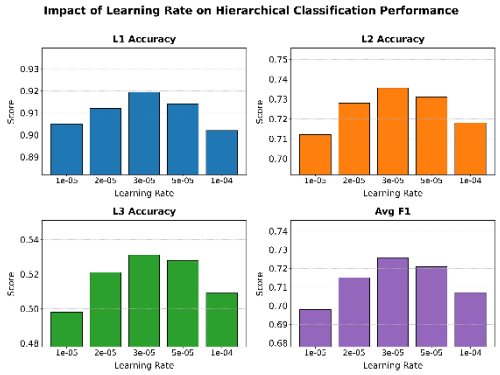

Figure 5. The impact of learning rate on hierarchical classification performance.

The results of the average F1 score provide even clearer evidence of this pattern. At the learning rate of 3e−5, the average F1 score reaches the highest value, demonstrating the model's ability to balance consistency across levels with fine-grained classification. This means that by selecting a proper learning rate, the model can not only improve overall accuracy but also enhance recognition of fine-grained categories while maintaining hierarchical consistency.

When the learning rate is further increased to 1e−4, the metrics show a decline. This suggests that an excessively large learning rate affects the stable update of parameters and weakens semantic modeling. In summary, the experiment verifies that learning rate, as a key hyperparameter, plays a decisive role in hierarchical text classification performance. Properly setting the learning rate maximizes the advantages of the BERT-based hierarchical classification method.

This paper also gives the impact of the number of encoding layers on representation ability, and the experimental results are shown in Figure 6.

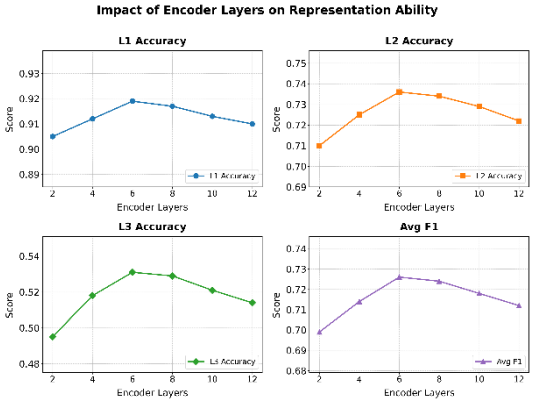

Figure 6. The impact of the number of coding layers on representation capabilitiesThe experimental results show that the number of encoder layers has a significant impact on the performance of hierarchical text classification.

 doi:10.20944/preprints202509.0750.v1

As the encoder depth increases from 2 to 6 layers, the model achieves consistent improvements across all metrics. This indicates that deeper structures provide stronger representational capacity for semantic modeling, which enhances both global and local feature learning in hierarchical classification.

For the first-level and second-level accuracy, the highest performance is achieved at 6 layers, reaching about 0.919 and 0.736, respectively. This suggests that an appropriate increase in encoder depth helps the model better capture the semantic features of higher-level labels and improves discriminative ability in middle-level label recognition. The results at this stage reflect the advantage of deeper networks in handling complex semantic relations. In hierarchical systems, semantic consistency can be maintained more effectively.

For the third-level accuracy and the average F1 score, a similar trend of "rising first and then declining" is observed as depth increases. At 6 layers, the model achieves the best performance, with third-level accuracy exceeding 0.53 and average F1 reaching 0.726. This shows that under an optimal configuration of depth, the model balances inter-class separability and intra-class compactness, leading to overall stable classification and reliable fine-grained prediction.

When the encoder depth further increases to 10 and 12 layers, the metrics begin to decline. This suggests that although deeper networks increase model capacity, they may also introduce parameter redundancy and overfitting, which weakens the effectiveness of representation. Overall, the experiment confirms that in hierarchical text classification tasks, properly controlling the number of encoder layers is crucial for enhancing semantic modeling ability and maintaining hierarchical consistency.

V. Conclusion

This study conducts a systematic analysis and empirical investigation of a BERT-based hierarchical classification method. By combining semantic representation with hierarchical constraints, the proposed model demonstrates strong stability and effectiveness across classification tasks at different levels. The experimental findings show that the method captures both global and local semantic information within multi-level label spaces, achieving significant improvements in first-level, second-level, and third-level accuracy as well as overall F1 score. These results not only confirm the potential of pre-trained language models in hierarchical text classification but also provide solid support for improving existing classification systems.

From a methodological perspective, this work integrates deep semantic modeling with hierarchical consistency regularization, addressing the limitations of traditional flat classification that overlook hierarchical label relations. The model maintains semantic integrity while alleviating the challenges posed by class imbalance and sparse lower-level labels, ensuring strong robustness even in long-tail categories and complex semantic scenarios. This approach offers a feasible solution for classification under multi-level label systems and lays a solid foundation for the design and optimization of future related tasks.

From the perspective of experimental analysis, the proposed model exhibits good adaptability under key hyperparameters such as learning rate and encoder depth. This further demonstrates its stability and applicability under different conditions. Consistent advantages are observed across accuracy, F1 score, t-SNE visualization, and confusion matrix analysis, confirming the strength of the method in both representation learning and hierarchical modeling. Such consistency across multiple evaluation dimensions provides important empirical support for the reliability and generalizability of the model.

Overall, this research contributes not only to the theoretical and methodological development of hierarchical text classification but also holds practical significance in multiple application domains. Tasks such as legal document classification, medical text processing, e-commerce product management, news topic recognition, and intelligent question answering can all benefit from the proposed hierarchical classification method. By improving fine-grained categorization and hierarchical consistency, these applications can achieve higher accuracy and better interpretability in

 doi:10.20944/preprints202509.0750.v1

large-scale information processing, thereby offering a stronger technical foundation for intelligent information services and knowledge management.

References

1. Hu J., Zhang B., Xu T., Yang H., and Gao M., "Structure-Aware Temporal Modeling for Chronic Disease Progression Prediction", arXiv preprint arXiv:2508.14942, 2025.
2. Wang X., Zhang X., and Wang X., "Deep Skin Lesion Segmentation with Transformer-CNN Fusion: Toward Intelligent Skin Cancer Analysis", arXiv preprint arXiv:2508.14509, 2025.
3. Zhang X. and Wang Q., "EEG Anomaly Detection Using Temporal Graph Attention for Clinical Applications", Journal of Computer Technology and Software, vol. 4, no. 7, 2025.
4. Su X., "Forecasting Asset Returns with Structured Text Factors and Dynamic Time Windows", Transactions on Computational and Scientific Methods, vol. 4, no. 6, 2024.
5. Wang Y., Sha Q., Feng H., and Bao Q., "Target-Oriented Causal Representation Learning for Robust Cross-Market Return Prediction", Journal of Computer Science and Software Applications, vol. 5, no. 5, 2025.
6. Du X., "Optimized convolutional neural network for intelligent financial statement anomaly detection", Journal of Computer Technology and Software, vol. 3, no. 9, 2024.
7. Xu Z., Liu X., Xu Q., Su X., Guo X., and Wang Y., "Time Series Forecasting with Attention-Augmented Recurrent Networks: A Financial Market Application", 2025.
8. Deng Y., "Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks", Journal of Computer Science and Software Applications, vol. 4, no. 6, 2024.
9. Gao D., "Graph Neural Recognition of Malicious User Patterns in Cloud Systems via Attention Optimization", Transactions on Computational and Scientific Methods, vol. 4, no. 12, 2024.
10. Chen H., Ma Q., Lin Z., et al., "Hierarchy-aware label semantics matching network for hierarchical text classification", Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4370-4379, 2021.
11. Kang T., Yang H., Dai L., Hu X., and Du J., "Privacy-Enhanced Federated Learning for Distributed Heterogeneous Data", 2025.
12. Wang Y., "Optimizing Distributed Computing Resources with Federated Learning: Task Scheduling and Communication Efficiency", Journal of Computer Technology and Software, vol. 4, no. 3, 2025.
13. Xing Y., Wang Y., and Zhu L., "Sequential Recommendation via Time-Aware and Multi-Channel Convolutional User Modeling", Transactions on Computational and Scientific Methods, vol. 5, no. 5, 2025.
14. Wei M., Xin H., Qi Y., Xing Y., Ren Y., and Yang T., "Analyzing data augmentation techniques for contrastive learning in recommender models", 2025.
15. Wang Z., Wang P., Huang L., et al., "Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification", arXiv preprint arXiv:2203.03825, 2022.
16. Jiang T., Wang D., Sun L., et al., "Exploiting global and local hierarchies for hierarchical text classification", arXiv preprint arXiv:2205.02613, 2022.
17. Zhan J., "Single-Device Human Activity Recognition Based on Spatiotemporal Feature Learning Networks", Transactions on Computational and Scientific Methods, vol. 5, no. 3, 2025.
18. Lou Y., "RT-DETR-Based Multimodal Detection with Modality Attention and Feature Alignment", Journal of Computer Technology and Software, vol. 3, no. 5, 2024.
19. Zi Y. and Deng X., "Joint Modeling of Medical Images and Clinical Text for Early Diabetes Risk Detection", Journal of Computer Technology and Software, vol. 4, no. 7, 2025.
20. Wang Z., Wang P., Liu T., et al., "HPT: Hierarchy-aware prompt tuning for hierarchical text classification", arXiv preprint arXiv:2204.13413, 2022.
21. Bao Q., "Advancing Corporate Financial Forecasting: The Role of LSTM and AI in Modern Accounting", Transactions on Computational and Scientific Methods, vol. 4, no. 6, 2024.
22. Tang T., Yao J., Wang Y., Sha Q., Feng H., and Xu Z., "Application of Deep Generative Models for Anomaly Detection in Complex Financial Transactions", Proceedings of the 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID), pp. 133-137, 2025.

 doi:10.20944/preprints202509.0750.v1

23. Xu Q. R., Xu W., Su X., Ma K., Sun W., and Qin Y., "Enhancing Systemic Risk Forecasting with Deep Attention Models in Financial Time Series", 2025.
24. Xu Z., Sheng Y., Bao Q., Du X., Guo X., and Liu Z., "BERT-Based Automatic Audit Report Generation and Compliance Analysis", Proceedings of the 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA), pp. 1233-1237, 2025.
25. Wang Y., "Entity-Aware Graph Neural Modeling for Structured Information Extraction in the Financial Domain", Transactions on Computational and Scientific Methods, vol. 4, no. 9, 2024.
26. Zi Y., Gong M., Xue Z., Zou Y., Qi N., and Deng Y., "Graph Neural Network and Transformer Integration for Unsupervised System Anomaly Discovery", arXiv preprint arXiv:2508.09401, 2025.
27. Du X., "Financial text analysis using 1D-CNN: Risk classification and auditing support", Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence, pp. 515-520, 2025.
28. Sheng Y., "Temporal Dependency Modeling in Loan Default Prediction with Hybrid LSTM-GRU Architecture", Transactions on Computational and Scientific Methods, vol. 4, no. 8, 2024.
29. Wang H., "Causal Discriminative Modeling for Robust Cloud Service Fault Detection", Journal of Computer Technology and Software, vol. 3, no. 7, 2024.
30. Zhu W., Wu Q., Tang T., Meng R., Chai S., and Quan X., "Graph Neural Network-Based Collaborative Perception for Adaptive Scheduling in Distributed Systems", arXiv preprint arXiv:2505.16248, 2025.
31. Sha Q., "Hybrid Deep Learning for Financial Volatility Forecasting: An LSTM-CNN-Transformer Model", Transactions on Computational and Scientific Methods, vol. 4, no. 11, 2024.
32. Ren Y., "Strategic Cache Allocation via Game-Aware Multi-Agent Reinforcement Learning", Transactions on Computational and Scientific Methods, vol. 4, no. 8, 2024.
33. Fang B. and Gao D., "Domain-Adversarial Transfer Learning for Fault Root Cause Identification in Cloud Computing Systems", arXiv preprint arXiv:2507.02233, 2025.
34. Cheng Y., "Selective Noise Injection and Feature Scoring for Unsupervised Request Anomaly Detection", Journal of Computer Technology and Software, vol. 3, no. 9, 2024.
35. Meng R., Wang H., Sun Y., Wu Q., Lian L., and Zhang R., "Behavioral Anomaly Detection in Distributed Systems via Federated Contrastive Learning", arXiv preprint arXiv:2506.19246, 2025.
36. Zhang R., "AI-Driven Multi-Agent Scheduling and Service Quality Optimization in Microservice Systems", Transactions on Computational and Scientific Methods, vol. 5, no. 8, 2025.
37. Yao G., Liu H., and Dai L., "Multi-Agent Reinforcement Learning for Adaptive Resource Orchestration in Cloud-Native Clusters", arXiv preprint arXiv:2508.10253, 2025.
38. Agrawal N., Kumar S., Bhatt P., et al., "Hierarchical text classification using contrastive learning informed path guided hierarchy", arXiv preprint arXiv:2506.04381, 2025.
39. Liu Y., Zhang K., Huang Z., et al., "Enhancing hierarchical text classification through knowledge graph integration", Findings of the Association for Computational Linguistics: ACL 2023, pp. 5797-5810, 2023.
40. Minaee S., Kalchbrenner N., Cambria E., et al., "Deep learning-based text classification: a comprehensive review", ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-40, 2021.
41. Zhou J., Ma C., Long D., et al., "Hierarchy-aware global model for hierarchical text classification", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1106-1117, 2020.
42. Plaud R., Labeau M., Saillenfest A., et al., "Revisiting hierarchical text classification: inference and metrics", arXiv preprint arXiv:2410.01305, 2024.
43. Wang X., "Time-Aware and Multi-Source Feature Fusion for Transformer-Based Medical Text Analysis", Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.
44. An T., et al., "A deep learning framework for boundary-aware semantic segmentation", Proceedings of the 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA), 2025.
45. Wu Q., "Internal Knowledge Adaptation in LLMs with Consistency-Constrained Dynamic Routing", Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.

 doi:10.20944/preprints202509.0750.v1

46. Wang X., Liu G., Zhu B., He J., Zheng H., and Zhang H., "Pre-trained Language Models and Few-shot Learning for Medical Entity Extraction", Proceedings of the 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA), pp. 1243-1247, 2025.
47. Quan X., "Layer-Wise Structural Mapping for Efficient Domain Transfer in Language Model Distillation", Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.
48. Zheng H., Zhu L., Cui W., Pan R., Yan X., and Xing Y., "Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models", 2025.
49. Lian L., Li Y., Han S., Meng R., Wang S., and Wang M., "Artificial Intelligence-Based Multiscale Temporal Modeling for Anomaly Detection in Cloud Services", arXiv preprint arXiv:2508.14503, 2025.
50. Zhang W., Tian Y., Meng X., Wang M., and Du J., "Knowledge Graph-Infused Fine-Tuning for Structured Reasoning in Large Language Models", arXiv preprint arXiv:2508.14427, 2025.
51. Yang T., Cheng Y., Ren Y., Lou Y., Wei M., and Xin H., "A Deep Learning Framework for Sequence Mining with Bidirectional LSTM and Multi-Scale Attention", Proceedings of the 2025 2nd International Conference on Innovation Management and Information System, pp. 472-476, 2025.
52. Xing Y., Yang T., Qi Y., Wei M., Cheng Y., and Xin H., "Structured Memory Mechanisms for Stable Context Representation in Large Language Models", arXiv preprint arXiv:2505.22921, 2025.
53. Qin Y., "Hierarchical semantic-structural encoding for compliance risk detection with LLMs", Transactions on Computational and Scientific Methods, vol. 4, no. 6, 2024.
54. Xu Q., "Unsupervised Temporal Encoding for Stock Price Prediction through Dual-Phase Learning", 2025.
55. Wang M., Kang T., Dai L., Yang H., Du J., and Liu C., "Scalable Multi-Party Collaborative Data Mining Based on Federated Learning", 2025.
56. Zhang R., Lian L., Qi Z., and Liu G., "Semantic and Structural Analysis of Implicit Biases in Large Language Models: An Interpretable Approach", arXiv preprint arXiv:2508.06155, 2025.
57. Wang Y., "Structured Compression of Large Language Models with Sensitivity-aware Pruning Mechanisms", Journal of Computer Technology and Software, vol. 3, no. 9, 2024.
58. Galke L., Scherp A., "Bag-of-words vs. graph vs. sequence in text classification: questioning the necessity of text-graphs and the surprising strength of a wide MLP", arXiv preprint arXiv:2109.03777, 2021.
59. Sahin U., Kucukkaya I. E., Toraman C., "ARC-NLP at PAN 2023: hierarchical long text classification for trigger detection", arXiv preprint arXiv:2307.14912, 2023.
60. Dai X., Chalkidis I., Darkner S., et al., "Revisiting transformer-based models for long document classification", arXiv preprint arXiv:2204.06683, 2022.
61. Im S. H., Kim G. B., Oh H. S., et al., "Hierarchical text classification as sub-hierarchy sequence generation", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12933-12941, 2023.