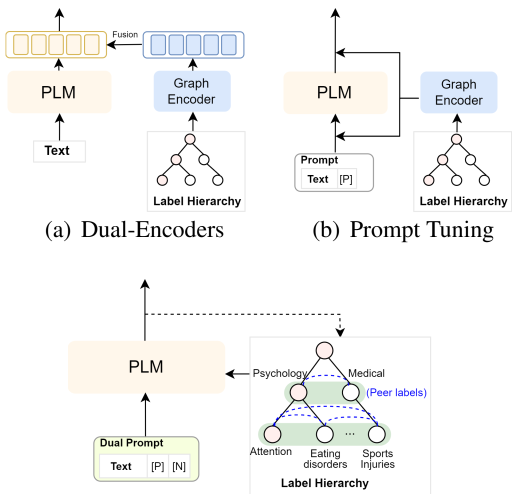
Dual Prompt Tuning based Contrastive Learning for Hierarchical Text Classification

Sishi Xiong[1], Yu Zhao[1], Jie Zhang[1],
Mengxiang Li[2], Zhongjiang He[1], Xuelong Li[1], Shuangyong Song[1]
[1]Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd
[2]China Telecom Corp Ltd
{xiongsishi, zhaoyu11, zhangj157}@chinatelecom.cn
limengx@126.com, hezj@chinatelecom.cn
xuelong_li@ieee.org, songshy@chinatelecom.cn

Abstract

Hierarchical text classification aims at categorizing texts into a multi-tiered tree-structured hierarchy of labels. Existing methods pay more attention to capture hierarchy-aware text feature by exploiting explicit parent-child relationships, while interactions between peer labels are rarely taken into account, resulting in severe label confusion within each layer. In this work, we propose a novel Dual Prompt Tuning (DPT) method, which emphasizes identifying discrimination among peer labels by performing contrastive learning on each hierarchical layer. We design an innovative hand-crafted prompt containing slots for both positive and negative label predictions to cooperate with contrastive learning. In addition, we introduce a label hierarchy self-sensing auxiliary task to ensure cross-layer label consistency. Extensive experiments demonstrate that DPT achieves significant improvements and outperforms the current state-of-the-art methods on BGC and RCV1-V2 benchmark datasets. [1]



(a) Dual-Encoders (b) Prompt Tuning

(c) Dual Prompt Tuning (Ours)

Figure 1: Architecture comparisons among existing methods and our proposed Dual Prompt Tuning.

1 Introduction

As a specialized sub-task of text classification, hierarchical text classification (HTC) has a wide range of applications in the realistic scenarios, such as intent recognition in dialogue system, commodity and book management (Cevahir and Murakami, 2016; Aly et al., 2019), where a large number of categories are organized into a tree-structured hierarchy. The ultimate goal of HTC is to categorize texts or documents from the topmost level to the finest granularity precisely. Due to the challenges of large-scale, imbalanced and complex label hierarchy (Mao et al., 2019), simply transferring flat multi-label text classification algorithms to deal with HTC tasks often fails to achieve sufficient performance.

Full use of the hierarchical structure of labels is the key to achieving well-performing classification in HTC tasks, which facilitates the model in predicting labels that align with hierarchical relationship. As shown in Figure 1, current mainstream methods (Zhou et al., 2020; Deng et al., 2021; Chen et al., 2021; Zhu et al., 2023) apply a dual-encoders framework to model the text and hierarchical structure separately, and then fuse them to obtain hierarchy-enhanced text features. Wang et al. (2022b) first proposes a hierarchy-aware prompt-tuning method, which incorporates the label hierarchy encoded by the Graph Attention Network into a soft prompt to bridge hierarchy and flat gap.

However, most studies pay close attention to exploit relations that explicitly displayed in the hierarchy, while internal interactions between "*peer labels*", which refer to a group of labels at the same hierarchical layer, are often neglected. PeerHTC (Song et al., 2023) recently tries to explore latent relevancy among peer labels with a complicated

[1]Code is publicly available at `https://github.com/ccx06/Dual-Prompt-Tuning-for-HTC`.

Findings of the Association for Computational Linguistics: ACL 2024, pages 12146–12158
August 11-16, 2024 ©2024 Association for Computational Linguistics

two-stage training procedure, in which peer and adjacent level-wise features are separately extracted by Graph Convolutional Neural Networks. Nevertheless, alleviating confusion within peer labels, especially fine-grained ones that share the same parent node at lower level, still remains challenging and highly valued.

To this end, we propose a novel **D**ual **P**rompt **T**uning (DPT) method, aimed at alleviating label confusion between peer labels. We put forward a Hierarchy-aware Peer-label Contrastive Learning (HierPCL) approach to extract discriminative pair-wise representations. In detail, we create an original dual prompt template containing both positive and negative label slots, and then perform label-wise contrastive learning on the embeddings of both two types of slots. Dual prompt is multi-functional, targeted for predicting positive labels and recognizing incorrect but confused negatives at each hierarchical layer. We furthermore incorporate a rank loss component into the contrastive loss function to enhance label consistency. Moreover, we design an adaptive hard negative sampling strategy and a hierarchy-injected label representation encoding method to further boost the performance of HierPCL.

Besides, we introduce a simple and effective label hierarchy self-sensing auxiliary task to keep our model in the best sense of holistic hierarchical structure. The basic idea is to internalize structural hierarchy knowledge to ensure the cross-layer consistency of the final path predictions. Instead of directly injecting label hierarchy into text semantics, we perform multi-task learning collaborating with label prediction to identify the consistency and correctness of each candidate label path.

The contributions are summarized as follows:

- We propose a Dual Prompt Tuning method for HTC tasks to address label confusion between peers at each hierarchical layer, magnifying the power of prompt.
- We put forward a Hierarchy-aware Peer-label Contrastive Learning approach based on the dual prompt, which contributes to obtaining aligned and discriminative features on prompt slots.
- We evaluate our proposed methods on four popular benchmark datasets against the strong baselines. Experimental results demonstrate the advantage of our proposal.

2 Related Work

2.1 Hierarchical Text Classification

The HTC algorithms can be generally divided into *local* and *global* approaches (Zangari et al., 2023). Local approaches (Wehrmann et al., 2018; Banerjee et al., 2019) construct multiple classifiers for different partitions of the label hierarchy usually in the "top-down" flow. Although there is a certain degree of connection between multiple classifiers, it is inevitable to lose the holistic structure information of label hierarchy. Global approaches use a single classifier to classify all labels with hierarchical dependencies simultaneously. Early works simplify the HTC task into a flat multi-label classification task, discarding the inherent hierarchical information in taxonomic labels. Later on, specialized hierarchy-aware methods are proposed. HiAGM (Zhou et al., 2020), HTCInfoMax (Deng et al., 2021), HiMatch (Chen et al., 2021), and HiTIN (Zhu et al., 2023) all employ the dual-encoders framework, which utilizes a text encoder and a structure encoder to learn the representations of texts and labels respectively, and then derives enhanced text embeddings based on both textual and structural information. Great progress has been made by leveraging advanced algorithms from other domains, i.e., applying sequence generative manners (Zhao et al., 2022; Ning et al., 2023; Huang et al., 2022) to mitigate label inconsistency phenomenon, data generation strategies (Wang et al., 2023) to enrich text diversity, contrastive learning methods (Wang et al., 2022a; Ji et al., 2023) to strengthen semantic expression, and prompt-tuning paradigm (Wang et al., 2022b; Ji et al., 2023) to exploit the potential of pre-trained language models (PLMs) (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2023; Liu et al., 2023b).

2.2 Prompt Tuning

Prompt Tuning (Schick and Schütze, 2021; Liu et al., 2023a) refers to the technique of tuning pre-trained language model by reconstructing downstream task into cloze test task, which bridges the gap in goals between fine-tuning and pre-training stages. It involves two main steps: (1) prompt template construction which generates a template containing special tokens, and (2) label word verbalizer design which defines a function from token embedding to answer words. There are two types of template construction methods. Hard prompt methods (Shin et al., 2020; Gao et al., 2021a; Han et al., 2022) directly concatenate explicit discrete tokens with the original text and maintain them unchanged throughout entire training process, which do not introduce any parameters. Soft prompt methods (Qin and Eisner, 2021; Lester et al., 2021; Gu et al., 2022; Liu et al., 2023c) convert prompt words into a group of continuous vectors as the template and update the vector based on specific contextual semantics and task objectives during training. The prompt-based method HPT (Wang et al., 2022b) adopts a soft prompt for HTC tasks, inserting a fixed number of learnable virtual label words into the input text. HierVerb (Ji et al., 2023) proposes a Multi-verbalizer (Multi-Verb) framework which integrates the hierarchical information, bringing significant performance improvement under few-shot settings.

2.3 Contrastive Learning

Contrastive learning (He et al., 2020; Chen et al., 2020) aims to pull anchor sample close to its positive samples while push it apart from negative samples, which has been proven to elevate the alignment and uniformity of feature space. Contrastive learning has various forms, differing in the construction of positive and negative pairs and loss formulas. Under self-supervised settings, positive samples are usually obtained through data augmentations or repeating dropout mask twice (Gao et al., 2021b) operation. Under supervised settings, positives are other samples of the same category (Khosla et al., 2020). Negative samples are selected from the remaining samples within a batch in self-supervised learning or samples belonging to distinct categories in supervised learning. Prototypical Contrastive Learning (Li et al., 2021) is proposed to enhance semantic discrimination and balance. It encourages instances to be closer to their assigned class prototypes, which can be established as the label semantics (Ma et al., 2022), the average embeddings of instances (Xiao et al., 2021) as well as learnable parameters (Cui et al., 2022).

3 Methodology

In this section, we present a detailed description of our proposed DPT model to address HTC tasks. As shown in Figure 2, our model is based on prompt-tuning framework with Multi-verbalizer. The principal innovations of DPT are twofold, including (1) the implementation of Hierarchy-aware Peer-label Contrastive Learning to obtain rich discriminative features, and (2) the incorporation of Label Hierarchy Self-sensing auxiliary task to enhance encoder's ability for an in-depth understanding of label hierarchy structure.

3.1 Preliminary

Given a set of inputs $D = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$ where $\mathbf{t}_i = \{x_j\}_{j=1}^{n}$ denotes a text composed of n words, and a predefined hierarchical label set $\mathcal{Y}$ which is commonly organized as a tree-like taxonomy structure $\mathcal{G}$, the goal of HTC is to select labels for $\mathbf{t}_i$ at each layer starting from the root label node of $\mathcal{G}$. Assuming L is the maximum depth of $\mathcal{G}$, the labels $\{y_1, y_2, ...\}$ of an input text correspond to single or multiple paths of the label tree, each of which typically consists of no more than L continuous individual labels with hierarchical relationship within $\mathcal{G}$.

3.2 Dual Prompt Tuning

Prompt template is utilized to wrap the original text to generate a new form of model input in prompt tuning paradigm. For example, the given text $\mathbf{t}_i$ is converted to "*[CLS] It was 1 level: [MASK] 2 level: [MASK] ... L level: [MASK].* $\mathbf{t}_i$ *[SEP]*" (Ji et al., 2023). Different from vanilla prompt, we exploit a dual prompt template to reserve two types of slot positions, instead of only insert positive label slots. For instance, a common dual prompt template is formulated as follow:

$$T = \{\text{[CLS] } \mathbf{t}_i \text{ [SEP] It belongs to [MASK] -...-[MASK] rather than [MASK]-...-[MASK][SEP]}\} \tag{1}$$

The number of *[MASK]* repetitions of positive or negative slots is equal to the depth of the label hierarchy L. In this paper, we define the *[MASK]* token inserted in the prompt template at the position of label slot as "*label mask token*". In the above example, positive label mask tokens locate between "It belongs to" and "rather than", while negative label mask tokens are behind "rather than". We'd like to emphasize that the number of prompt words introduced by dual prompt template is linearly related to the maximum depth of label hierarchy rather than the number of full labels.

Consistent with other competitive methods, we employ BERT (Devlin et al., 2019) as the backbone of our model to encode input texts and obtain all

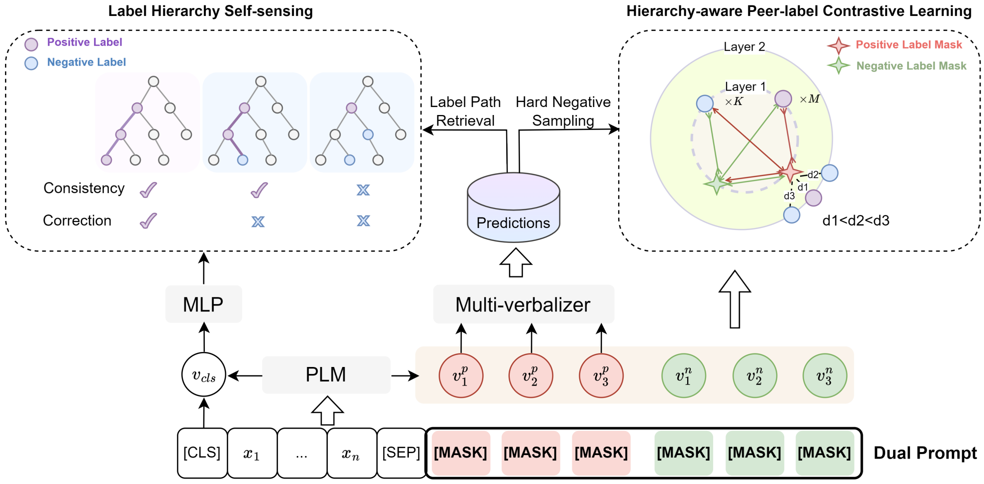


Figure 2: Model architecture of Dual Prompt Tuning (DPT). It consists of two innovative modules: Hierarchy-aware Peer-label Contrastive Learning and Label Hierarchy Self-sensing Task. Label mask token embeddings and label representations are encoded in the unified embedding space by the PLM, and contrastive loss is calculated according to their affiliation and hierarchical relationship. Label Hierarchy Self-sensing task is used to simultaneously restrain path predictions with correctness and consistency. Note that the figure only depicts the Hierarchy-aware Peer-label Contrastive learning between the first and second levels.

token embeddings:

$$V = \text{BERT}(T(\mathbf{t}_i)) \tag{2}$$

Let v_l^p and v_l^n respectively represent the embeddings of positive and negative label mask token at the l-th layer. We inherit HierVerb (Ji et al., 2023) to adopt a depth-oriented Multi-verbalizer framework mapping label mask token embeddings $\{v_l^p\}_{l=1}^L$ to label words. Probability distribution of $\mathbf{t}_i$ can be expressed as:

$$\begin{aligned} Z &= \{z_l\}_{l=1}^L \\ &= \{Ver_1(v_1^p), ..., Ver_L(v_L^p)\} \end{aligned} \tag{3}$$

where the l-th verbalizer Ver_l acts in predicting l-level labels. More details about Multi-verbalizer framework can be found in Ji et al. (2023).

3.3 Hierarchy-aware Peer-label Contrastive Learning

To extract hierarchy-aware dicriminative feature on the basic of dual prompt tuning, the ideal embeddings at label mask tokens should satisfy the following intents: (1) Token embeddings of the positive label mask tokens should be close to representations of their positive labels, and far away from negative labels in the feature space. The same desire applies to token embeddings of negative label mask tokens. (2) The semantic similarity between high-level label and its ground truth sub-label is expected to be greater than that with other sub-labels, further greater than that with sub-labels of other nodes at the same hierarchical layer. Based on the above, we propose a Hierarchy-aware Peer-label Contrastive Learning (HierPCL) method to capture latent semantic relevancy between peer labels as well as parent-child labels.

Objective of HierPCL The basic idea of HierPCL is to encourage the embeddings of the label mask tokens encoded by PLM closer to the representations of their positive labels which should be filled in the label slots of the template. The objective function of HierPCL consists of three components: positive label contrastive learning, negative label contrastive learning and cross-hierarchical rank loss.

(1) Positive label contrastive learning is performed on positive label mask tokens. The target positives are M ground truth labels while the negatives are the sampled K negative labels and negative label mask token. The loss function is formulated as:

$$\mathcal{L}_{CL}^p = -\frac{1}{L}\sum_{l=1}^{L} \log \frac{\sum_{m=1}^{M} \exp(s(v_l^p, r_{l,m}^p)/\tau)}{\sum_{u \in \mathcal{A}_l^p} \exp(s(v_l^p, u)/\tau)} \tag{4}$$

where $r_{l,m}^p$ and $r_{l,k}^n$ respectively denote the representation vectors of the m-th positive labels and the k-th negative label in the l-th layer of sentence $\mathbf{t}_i$. $s(\cdot)$ represents cosine similarity function, and τ is the temperature coefficient. All participants above are denoted as $\mathcal{A}_l^p := \{\{r_{l,m}^p\}_{m=1}^M, \{r_{l,k}^n\}_{k=1}^K, v_l^n\}$, including representation vectors of positive and negative labels, and embeddings on negative label mask token at the same hierarchical layer.

(2) Negative label contrastive learning is performed on negative label mask tokens. Opposite to $\mathcal{L}_{LC}^{p}$, the target positives are negative labels of this instance while the negatives are ground truth labels and positive label mask token. Let $\mathcal{A}_l^n := \{\{r_{l,m}^p\}_{m=1}^{M}, \{r_{l,k}^n\}_{k=1}^{K}, v_l^p\}$, the loss function is formulated as:

$$\mathcal{L}_{CL}^{n} = -\frac{1}{L}\sum_{l=1}^{L}\log\frac{\sum_{k=1}^{K}\exp(s(v_l^n, r_{l,k}^n)/\tau)}{\sum_{u\in\mathcal{A}_l^n}\exp(s(v_l^n, u)/\tau)} \quad (5)$$

(3) Cross-hierarchical rank loss aims to align label token embeddings with the representations of their sub-labels. In other words, high-level labels tend to have higher semantic similarities with positive child-labels compared to negative child-labels and other peer labels. The loss function is formulated as:

$$\begin{aligned}\mathcal{L}_R = \sum_{l=1}^{L-1}\Big(&\sum_{m=1}^{M}\sum_{k=1}^{K}\max(0, s(v_l^p, r_{l+1,k}^n) - s(v_l^p, r_{l+1,m}^p)) \\ &+ \sum_{\bar{k}}\sum_{\hat{k}}\max(0, s(v_l^p, r_{l+1,\hat{k}}^n) - s(v_l^p, r_{l+1,\bar{k}}^n))\Big)\end{aligned} \quad (6)$$

where $r_{l+1,\bar{k}}^n$ denotes the representation vector of the negative label at $(l+1)$-th layer whose parent label is positive at layer l, while $r_{l+1,\hat{k}}^n$ denotes that of other negative label that does not belong to a positive label at the next higher level.

Finally, the loss function of HierPCL is formulated as follow:

$$\mathcal{L}_1 = \alpha\mathcal{L}_{CL}^p + (1-\alpha)\mathcal{L}_{CL}^n + \beta\mathcal{L}_R \quad (7)$$

where α and β are hyper-parameters used for balancing relative weights of the three components.

Top-K Hard Negative Sampling The selection strategy of negative samples is critical to contrastive learning. Since the estimation of label predictions can be considered as a reliable source for generating hard negatives, we devise an adaptive self-produced hard negative sampling strategy under the guidance of the model's own predictions during training. It adopts the top K hard negative labels according to confidence scores ranked in descending order at each hierarchical layer. In our experiments, K is set to 10% of the number of labels at each layer, which achieves a balance between performance and memory consumption. The effects of different settings of K are described in Appendix B in detail.

Hierarchy-injected Label Representation Label representation vectors are learned through the shared PLM without prior statistics. To avoid the side impact of the overlapping interaction between label names and text words, we assign an unique fabricated symbol for each label, like "*L0*", and add these symbols to the vocabulary list used in model training. To incorporate hierarchy information to label representation, we flatten the parent-child hierarchy of a label to form a label sequence, as follow:

$$Q = \{\text{[CLS]}\ W\ \text{[SEP]}\ W^f\ \text{[SEP]}\ \{W^c\}\ \text{[SEP]}\} \quad (8)$$

Assuming W is the symbol of current label, then W^f means the parent label symbol of current label W, and $\{W^c\}$ is on behalf of symbols of all child labels. "Root" and "None" are used as fictitious tokens when the parent label or child label doesn't exist. We use the embedding on W enriched with hierarchical dependencies as the label representation r.

3.4 Label Hierarchy Self-sensing Task

For the purpose of elevating the model's perceptual ability of holistic hierarchical structure, we introduce a Label Hierarchy Self-sensing auxiliary task, consisting of two sub-tasks: (1) determining whether the predicted label nodes at each level can form valid label paths, and (2) determining whether the predicted label paths are completely correct. A simple base unit of feed-forward neural network is utilized on the top of *[CLS]* token following PLM to execute the auxiliary task. The consistency and correction loss functions are respectively defined based on the Binary Cross Entropy (BCE) (De Boer et al., 2005) function:

$$\mathcal{L}_{con} = \text{BCE}(\bar{y}_{con}, \bar{p}_{con}) \quad (9)$$

$$\mathcal{L}_{cor} = \text{BCE}(\bar{y}_{cor}, \bar{p}_{cor}) \quad (10)$$

where $\bar{p}_{con}$ represents the probability of that the predictions can fully correspond to label paths in the label tree, and $\bar{p}_{cor}$ represents the probability of that predictions are all right. We retrieve all label paths from the predicted label nodes output by verbalizers. $\bar{y}_{con} = 1$ if all predicted label nodes can exactly form label paths. $\bar{y}_{cor} = 1$ if all combined label paths are the positive labels of the sentence. Otherwise, the value of $\bar{y}_{con}$ or $\bar{y}_{cor}$ is 0.

Finally, the loss function of the auxiliary task is formulated as:

$$\mathcal{L}_2 = \mathcal{L}_{con} + \mathcal{L}_{cor} \tag{11}$$

3.5 Multi-task Training

Overall, multi-task training objective is to minimize the weighted combination of classification loss, peer-label contrastive learning loss, label hierarchy self-sensing loss and MLM loss retaining from BERT pre-training. The classification loss function can be contingent upon specific circumstances. For the sake of universality, the standard Binary Cross-Entropy function is employed:

$$\mathcal{L}_{CLS} = \sum_{l=1}^{L} \text{BCE}(y_l, z_l) \tag{12}$$

Final joint loss can be formulated as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{MLM} + \mathcal{L}_{CLS} \\ &+ \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 \end{aligned} \tag{13}$$

where λ_1 and λ_2 are hyper-parameters.

4 Experiment

4.1 Experiment Setup

Datasets We conduct experiments on 4 benchmark datasets: Web-of-Science (WoS) (Kowsari et al., 2017), NYTimes (NYT) (Sandhaus, 2008), RCV1-V2 (Lewis et al., 2004) and Blurb Genre Collection (BGC) [2] (Aly et al., 2019) . The label taxonomy of WoS is single-path whereas the remaining three datasets are for multi-path HTC. The detailed statistics of these datasets are listed in Table 1.

Evaluation Metrics The performance of our methods is evaluated by popular Micro-F1 and Macro-F1 metrics. In addition, path-constrained C-MicroF1 and C-MacroF1 metrics proposed by Yu et al. (2022) are used to measure the hierarchical path consistency for comprehensive evaluation, in which an output label is considered as correct only when all its ancestor nodes are accurately predicted.

Implementation Details The backbone of our model is initialized with bert-base-uncased[3]. The batch size is set to 32 for BGC and 16 for other datasets. The AdamW optimizer is used with the learning rate of 2e-5 for WoS and 3e-5 for others. We apply early stopping strategy with 5 patient epochs. For fair comparison, we perform the same data processing and spliting methods as HPT. The reported results of our main experiments are the average score of 5 runs over different random seeds. Experimental settings of hyper-parameters are described in Appendix A.

Dataset	WoS	RCV1-V2	BGC	NYT
L	2	4	4	8
$\lvert\mathcal{Y}\rvert$	141	103	146	166
Avg($\lvert\mathcal{Y}_i\rvert$)	2.0	3.24	3.01	7.6
#Train	30070	20833	58715	23345
#Val	7518	2316	14785	5834
#Test	9397	781265	18394	7292

Table 1: Datasets statistics. L, $\lvert\mathcal{Y}\rvert$ and Avg($\lvert\mathcal{Y}_i\rvert$) represent the maximum depth, total number of categories and the average number of labels per sample, respectively.

Baselines We compare our methods with the following advanced HTC methods:

- **HiAGM**, **HTCInfoMax** and **HiMatch** are dual-encoders based classic methods. They derive joint embeddings on text and labels from interactive fusion or matching mechanisms.

- **HGCLR** incorporates label hierarchy into text encoder through hierarchy-guided contrastive learning between texts and their generated positive samples with the most closest label paths.

- **Seq2Tree** and **PAAM-HiA-T5** treat HTC as a sequence generation task. Seq2Tree designs a constrained decoding strategy with dynamic vocabulary to ensure label consistency. PAAM-HiA-T5 proposes a multi-level sequential label generative T5 model with a path-adaptive attention mechanism to focus on label dependency prediction.

- **HPT** exploits the effects of prompt-tuning by a dynamic virtual template and a zero-bounded multi-label cross entropy loss. It achieves the current state-of-the-art performance on most datasets.

- **HiTIN** introduces the structural entropy to construct a coding tree for the label hierarchy and then builds a novel structure encoder to enhance text representations.

[2] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/blurb-genre-collection.html

[3] https://huggingface.co/bert-base-uncased

Model	WoS		RCV1-V2		BGC		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BERT(Wang et al., 2022a)	85.63	79.07	85.65	67.02	-	-	78.24	66.08
BERT+HiMatch(Chen et al., 2021)	86.70	81.06	86.33	68.66	78.89	63.19	76.79	63.89
BERT+HiAGM(Wang et al., 2022a)	86.04	80.19	85.58	67.93	-	-	78.64	66.76
BERT+HTCInfoMax(Wang et al., 2022a)	86.30	79.97	85.53	67.09	-	-	78.75	67.31
HGCLR(Wang et al., 2022a)	87.11	81.20	86.49	68.31	-	-	78.86	67.96
Seq2Tree(Yu et al., 2022)	87.20	**82.50**	86.88	70.01	79.72	63.96	-	-
PAAM-HiA-T5(Huang et al., 2022)	**90.36**	81.64	87.22	70.02	-	-	77.52	65.97
HPT(Wang et al., 2022b)	87.16	81.93	87.26	69.53	81.32†	66.69†	80.42	**70.42**
HiTIN(Zhu et al., 2023)	87.19	81.57	86.71	69.95	-	-	79.65	69.31
DPT (Ours)	87.25	81.51	**87.76**	**70.78**	**81.85**	**68.21**	**80.56**	70.28

Table 2: Experimental results on four HTC datasets. The best results are in bold format, and the second-best results are in underlined format. The results of BERT+HiMatch on NYT dataset are reported by Huang et al. (2022). The results of BERT+HiMatch on BGC are reported by Yu et al. (2022). † means the results are reproduced upon the release project by ourselves.

4.2 Main Result

Experimental results are shown in Tabel 2. Our model consistently outperforms previous advanced approaches across 3 datasets except for WoS. On WoS dataset with label depth of 2, our proposed DPT achieves comparable results with HPT but decreased performance compared to PAAM-HiA-T5 model. Our model establishes the state-of-the-art results on RCV1-V2 and BGC datasets. It improves 0.5% and 0.76% absolute Micro-F1 and Macro-F1 on RCV1-V2 dataset compared to the current best results. The performance boost of Micro-F1 and Macro-F1 on BGC reaches 0.53% and 1.52% over HPT model. The significant improvements on Macro-F1 metric suggest that our model excels in identifying sparse labels. On NYT dataset, our model surpasses HPT on Micro-F1 by 0.14% but slightly lower on Macro-F1.

Without introducing any additional network parameters to extract the semantics of labels and their hierarchical structures, our model surprisingly outperforms previous methods of encoding label names and label hierarchies using GNNs. Compared to HGCLR, which employs instance-level contrastive learning with complex positive sample generation operation, DPT achieves performance gains across all datasets.

4.3 Results on Label Consistency

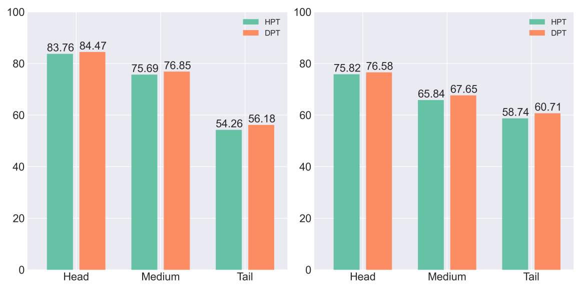
For HTC tasks, cross-layer label consistency is an essential evaluation factor, which signifies that multiple labels predicted by the model at each layer should conform to the predefined hierarchical structure. Table 3 illustrates the label consistency performance of our proposed DPT and the state-of-the-art HPT model. DPT improves consistency of label hierarchy on RCV1-V2 and BGC by a large margin, respectively exceeding HPT by 1.49% and 1.61% on C-MacroF1 metric. Although our model focuses more on the interaction between peer labels at the same layer, the knowledge of label hierarchy has also been internalized. The accuracy of both individual label nodes and united label paths has been improved, indicating that our methods are reasonable and efficient.

Model	RCV1-V2		BGC		NYT	
	C-MiF1	C-MaF1	C-MiF1	C-MaF1	C-MiF1	C-MaF1
HPT	86.80	68.71	80.88	65.36	79.33	**68.80**
DPT (Ours)	**87.47**	**70.20**	**81.43**	**66.97**	**79.77**	68.70

Table 3: Evaluation results of label consistency. C-MiF1 and C-MaF1 are the abbreviations for C-MicroF1 and C-MacroF1.

4.4 Results on Imbalanced Hierarchy

To further clarify the superiority of our methods, we explore the performance on the imbalanced hierarchy. Following long-tailed learning setting, we sort the training set in descending order based on the quantity of class instances and evenly cluster all categories into head, middle, and tail groups. The visualization results on Macro-F1 metric of different groups are shown in Figure 3. It's evident that the performance of DPT is better than that of HPT on imbalanced hierarchy in RCV1-V2 and BGC datasets. DPT shows significant improvements on tail classes with few training samples, demonstrating the effectiveness of our methods in eliminating the impact resulting from imbalanced distribution.



(a) RCV1-V2 (b) BGC

Figure 3: Macro-F1 score on head, medium and tail class groups.

Ablation Models	BGC			
	MiF1	MaF1	C-MiF1	C-MaF1
Multi-Verb(Baseline)	81.38	66.74	80.92	65.58
DPT(Ours)	**81.85**	**68.21**	**81.43**	**66.97**
r.m. $\mathcal{L}_2$	81.71	68.09	81.27	66.56
r.m. $\mathcal{L}_{LC}^{n}$	81.62	67.35	81.35	66.11
r.m. $\mathcal{L}_R$	81.83	67.99	81.40	66.50
r.m. $\mathcal{L}_R$ & $\mathcal{L}_{LC}^{n}$	81.36	67.41	80.97	66.26
r.p. Random Sampling	81.48	67.46	81.03	66.59

Table 4: Ablation study on BGC dataset.

4.5 Ablation Study

To investigate the effects of each component of our proposed model, we implement different variants and conduct experiments on BGC and RCV1-V2 dataset. The Results are shown in Table 4 and Tabel 5 respectively. Upon only employing the HierPCL module, the performance in all metrics realizes considerable enhancement and are superior to the current state-of-the-art models, confirming its significant effectiveness and reliability. By removing negative contrastive part, the scores undergo sharp declines, which demonstrates that negative label contrastive learning plays a prominent role in HierPCL. As a strong contrast, we replace our self-produced hard negative sampling with random sampling, resulting in decrease in metrics, which validates the advantage of our negative sampling strategy. The cross-hierarchical rank loss in HierPCL and label hierarchy self-sensing auxiliary task are evidenced to improve the model's performance, specifically with regard to the Macro-F1 metric.

4.6 Insight into Case Effects

In order to gain insight into practical effects of our model, we conduct detailed case studies on the test set. We define 3 types of label errors from the perspective of multi-label classification, as follows:

Ablation Models	RCV1-V2			
	MiF1	MaF1	C-MiF1	C-MaF1
Multi-Verb(Baseline)	87.19	69.16	86.68	68.31
DPT(Ours)	**87.76**	**70.78**	**87.47**	**70.20**
r.m. $\mathcal{L}_2$	87.34	70.28	87.01	69.45
r.m. $\mathcal{L}_{LC}^{n}$	87.14	70.05	86.61	69.32
r.m. $\mathcal{L}_R$	87.47	70.35	87.11	69.52
r.m. $\mathcal{L}_R$ & $\mathcal{L}_{LC}^{n}$	87.22	69.52	86.79	68.73
r.p. Random Sampling	87.26	69.55	86.94	68.93

Table 5: Ablation study on RCV1-V2 dataset.

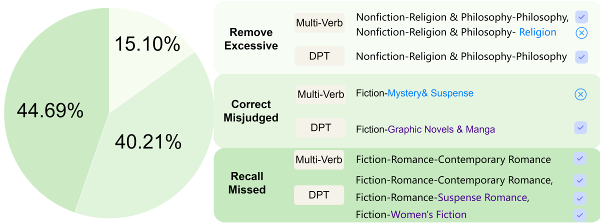


Figure 4: Three types of errors corrected by our model.

- Excessive, refers to the label which is unnecessarily identified as one of the positive labels for the instance.
- Misjudged, refers to the label which is mistakenly identified as a positive label while the instance actually belongs to another peer label.
- Missed, refers to the label that the model has failed to recall.

We separately calculate the distribution of label error types for baseline model (typical prompt-tuning with Multi-Verb framework) and our improved model on the test set. As shown in Figure 4, we find that the optimization effects of our model are manifested in recalling missed labels, correcting misjudged labels and removing excessive labels, respectively accounting for 44.69%, 40.21% and 15.10% of the proportion. It demonstrates the strong power of DPT to capture discriminative representation and further relieve label confusion. Some specific cases are illustrated in Appendix D.

4.7 Analysis of Computational Complexity

We conduct experiments with an NVIDIA Tesla V100 on time efficiency and computational resources on RCV1-V2 dataset. The quantitative results are listed in Appendix C. Compared to prompt-based HPT model, our approach necessitates longer training time, primarily attributed to contrastive learning, which inherently possesses a high time complexity and slightly elevates computational resources requirements. Despite all this, both factors remain within acceptable range. Furthermore, label-wise contrastive learning has superior time efficiency than instance-level contrastive learning. We devise an effective negative sampling strategy to reduce computational costs. When performing model inference, our methods do not involve any additional time-consuming steps, such as loss calculations, and exhibit faster speed compared to models that employ GNN as the structure encoder.

5 Conclusion

In this paper, we present a novel Dual Prompt Tuning method for HTC tasks. Firstly, we propose a Hierarchy-aware Peer-label Contrastive Learning approach to alleviate confusion between peer labels. An original dual prompt template is created with slots for both positive and negative label, on which contrastive learning is performed at each layer. Secondly, to further strengthen knowledge understanding of label hierarchy structure, we design a Label Hierarchy Self-sensing auxiliary task to identify consistency and correctness of model predictions. Experimental results illustrate that our proposed DPT method achieves significant improvements on popular HTC datasets. It excels in precisely recognizing negative labels and contributes to obtaining hierarchy-aware discriminative features. Notably, DPT exhibits remarkable effectiveness in preserving label path consistency and addressing imbalanced hierarchy challenge.

6 Limitations

In our work, the applicability of DPT is restricted by the scale of hierarchical labels, considering the following two factors: (1) The sequence length of input text for PLMs is constrained to a maximum value, and adding prompts consumes some tokens. However, the number of prompt words introduced by DPT is only linearly related to the depth of the label tree rather than the number of full labels, which has minimal impact. (2) As outlined in Section 3.3, when computing label representation vector, the maximum allowed number of child labels $\{W^c\}$ is close to the maximum sequence length of the PLM. Although DPT applies to most scenarios, it is recognized that our methods may require further improvement to accommodate huge-scale hierarchical label systems.

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ali Cevahir and Koji Murakami. 2016. Large-scale multi-class and hierarchical product categorization for an E-commerce giant. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 525–535, Osaka, Japan. The COLING 2016 Organizing Committee.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *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)*, pages 4370–4379, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland. Association for Computational Linguistics.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip Yu. 2021. HTCInfoMax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3259–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Wei Huang, Chen Liu, Bo Xiao, Yihua Zhao, Zhaoming Pan, Zhimin Zhang, Xinyun Yang, and Guiquan Liu. 2022. Exploring label hierarchy in a generative way for hierarchical text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1116–1127, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2918–2933, Toronto, Canada. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. 5:361–397.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023b. icsberts: Optimizing pre-trained language models in intelligent customer service. *Procedia Computer Science*, 222:127–136. International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023).

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023c. Gpt understands, too. *AI Open*.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 445–455, Hong Kong, China. Association for Computational Linguistics.

Bo Ning, Deji Zhao, Xinjian Zhang, Chao Wang, and Shuangyong Song. 2023. Ump-mg: A uni-directed message-passing multi-label generation model for hierarchical text classification. *Data Science and Engineering*, 8(2):112–123.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Evan. Sandhaus. 2008. The new york times annotated corpus ldc2008t19. *Philadelphia: Linguistic Data Consortium*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Junru Song, Feifei Wang, and Yang Yang. 2023. Peer-label assisted hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3758, Toronto, Canada. Association for Computational Linguistics.

Yue Wang, Dan Qiao, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2023. Towards better hierarchical text classification with data generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7722–7739, Toronto, Canada. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022a. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3740–3751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. 2018. Hierarchical multi-label classification networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084. PMLR.

Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does head label help for long-tailed multi-label text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:14103–14111.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1865–1869.

Alessandro Zangari, Matteo Marcuzzo, Michele Schiavinato, Matteo Rizzo, Andrea Gasparetto, Andrea Albarelli, et al. 2023. Hierarchical text classification: a review of current research. *EXPERT SYSTEMS WITH APPLICATIONS*, 224.

Deji Zhao, Bo Ning, Shuangyong Song, Chao Wang, Xiangyan Chen, Xiaoguang Yu, and Bo Zou. 2022. Tosa: A top-down tree structure awareness model for hierarchical text classification. In *Web and Big Data: 6th International Joint Conference, APWeb-WAIM 2022, Nanjing, China, November 25–27, 2022, Proceedings, Part II*, page 23–37, Berlin, Heidelberg. Springer-Verlag.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

He Zhu, Chong Zhang, Junjie Huang, Junran Wu, and Ke Xu. 2023. HiTIN: Hierarchy-aware tree isomorphism network for hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7809–7821, Toronto, Canada. Association for Computational Linguistics.

A Hyper-parameter Settings

We list the hyper-parameter settings of four datasets in Table 6 for reproducibility.

Parameter	WoS	RCV1-V2	BGC	NYT
α	0.2	0.6	0.6	0.2
β	0.1	0.2	0.1	0.1
λ_1	0.5	1.0	0.5	0.1
λ_2	0.6	0.5	0.2	0.2

Table 6: Hyper-parameter settings.

B Performance of Different Negative Sampling Ratio

To elaborate on the rationality behind selecting a K value as 10% of the label count at each layer, we compare the performance when the negative sampling rate is set at 10% and 100%. From Table 7 and Table 8, it's obvious that 10% negative labels are sufficient, retaining the vast majority of accuracy on RCV1-V2 and even surpassing the performance of using all negatives on BGC dataset. It indicates that blindly increasing the number of negative samples is not always effective. We will explore the impact of positive and negative label ratio in the future work.

Ratio	MiF1	MaF1	C-MiF1	C-MaF1
0.1	87.76	**70.78**	87.47	**70.20**
1	**87.81**	70.01	**87.50**	69.24

Table 7: Results of different negative sampling ratio on RCV1-V2 dataset.

Ratio	MiF1	MaF1	C-MiF1	C-MaF1
0.1	**81.85**	**68.21**	**81.43**	**66.97**
1	81.75	67.95	81.41	66.62

Table 8: Results of different negative sampling ratio on BGC dataset.

C Computational Complexity of DPT

Table 9 compares computational complexity between our DPT model and the current state-of-the-art prompt-based HPT model on RCV1-V2 dataset.The results show that our model has faster inference speed and acceptable training computational costs.

Model	#Params (M)	Training time (min/epoch)	Training memory usage(G)	Inference (ms/sample)
HPT	114.92	16.06	24.8	23.1
DPT(Ours)	112.67	28.19	26.9	20.7

Table 9: Computational complexities of HPT and DPT.

D Case Study

DPT performs Peer-label Contrastive Learning at each level, which enhances the model's representation abilities. Introduction of cross-hierarchical rank loss and label hierarchy self-sensing auxiliary task improves label consistency. To look into the practical effects, we conduct adequate case studies on the BGC dataset. Compared to the strong baseline HPT model, improvements of DPT are mainly reflected in recalling missed labels, correcting misjudged labels, removing excessive labels, and further correcting label inconsistencies. Table 10 provides some examples. Note that labels predicted by DPT model in the Table 10 are consistent with the ground truth labels.

Text	Secret, Silent Screams: For fans of Gillian Flynn, Caroline Cooney, and R.L. Stine comes Secret, Silent Screams from four-time Edgar Allen Poe Young Adult Mystery Award winner Joan Lowery Nixon. Is Barry's death the latest tragedy in a string of suicides at Farrington Park High School? Or is it murder? Marti is sure her friend Barry didn't take his own life, but no one will believe her except Police Officer Prescott. But opening an investigation takes time, and Marti is determined to find her friend's killer soon. Because even now he could be planning his next crime... "Enthralling suspense...satisfying[,]...[and an] intricate plot." –Publishers WeeklyFrom the Paperback edition.
DPT (Ours)	Teen & Young Adult-Teen & Young Adult Mystery & Suspense, Teen & Young Adult-Teen & Young Adult Fiction, Teen & Young Adult-Teen & Young Adult Social Issues
Multi-Verb (Baseline)	Teen & Young Adult-Teen & Young Adult Mystery & Suspense, Teen & Young Adult-Teen & Young Adult Fiction
Text	Summer in Eclipse Bay: A special message from Jayne Ann KrentzDear Reader:Summer has arrived in Eclipse Bay and things are definitely heating up between the Hartes and the Madisons. It seems that the mysterious new gallery owner, Octavia Brightwell, is thinking about having a scandalous fling with that rogue Nick Harte before she leaves town. As far as Nick is concerned, a short-term affair sounds perfect. But it isn't going to be easy.One big obstacle is Mitchell Madison. For reasons of his own, Mitchell has taken it upon himself to play guardian to Octavia. He's made it clear that if Nick fools around with her, there will be a price to pay. And then there's Nick's young son, Carson, who has his own agenda where Octavia is concerned. He doesn't want his father messing up his plans.Summer in Eclipse Bay is going to be eventful this year. Some long-buried secrets from the infamous Harte-Madison feud are about to surface. The past and the present are on a collision course. I hope you'll join me to watch the fireworks.Happy reading . . .Jayne Ann Krentz.
DPT (Ours)	Fiction-Romance-Contemporary Romance, Fiction-Romance-Suspense Romance, Fiction-Women's Fiction
Multi-Verb (baseline)	Fiction-Romance-Contemporary Romance
Text	Corto Maltese: Beyond The Windy Isles: The second volume in the definitive English language edition of Hugo Pratt's masterpiece, Corto Maltese, presented in the original oversized B&W format and with new translations made from Pratt's original Italian scripts. "Mushroom Heads" begins in Maracaibo, Venezuela, where Corto Maltese and Professor Steiner lead an expedition on the trail of the legendary El Dorado, financed by the antiquarian Levi Colombia. In "Banana Conga," Corto has his first and nearly fatal encounter with the beautiful yet dangerous mercenary Venexiana Stevenson. Within this framework of adventure, Hugo Pratt weaves themes dealing with the exploitation of indigenous people, the noble struggle to gain freedom and independence, and how cowardice can poison men of all classes. The action, set in 1917, takes Corto Maltese from the Mosquito Coast to Barbados to a deadly struggle among Jivaro head-hunters in the Peruvian Amazon
DPT (Ours)	Fiction-Graphic Novels & Manga
Multi-Verb (Baseline)	Fiction-Mystery & Suspense
Text	Nappily Ever After: SOON TO BE A NETFLIX ORIGINAL FILM STARRING SANAA LATHANWhat happens when you toss tradition out the window and really start living for yourself? Venus Johnston has a great job, a beautiful home, and a loving live-in boyfriend named Clint, who happens to be a drop-dead gorgeous doctor. She also has a weekly beauty-parlor date with Tina, who keeps Venus's long, processed hair slick and straight. But when Clint–who's been reluctant to commit over the past four years–brings home a puppy instead of an engagement ring, Venus decides to give it all up. She trades in her long hair for a dramatically short, natural cut and sends Clint packing. It's a bold declaration of independence–one that has effects she never could have imagined. Reactions from friends and coworkers range from concern to contempt to outright condemnation. And when Clint moves on and starts dating a voluptuous, long-haired beauty, Venus is forced to question what she really wants out of life. With wit, resilience, and a lot of determination, she finally learns what true happiness is–on her own terms. Told with style, savvy, and humor, Nappily Ever After is a novel that marks the debut of a fresh new voice in fiction.
DPTL (Ours)	Fiction-Women's Fiction
Multi-Verb (Baseline)	Fiction-Romance
Text	Malice: Despite our tendencies to separate the mind and body, good and evil, Flahault argues that both stem from the same source within us. This knot, inherent to the human condition, is the tension between our desire for absolute self-affirmation and the fact that each of us can only exist through mediation by others. The dependence on others weighs heavy on our shoulders, hampering our very existence.Malice, then, is not merely a result of our biological constitution, but is also a response to our feelings. These can often resemble those of Milton's and Shelley's monsters, stories the author calls upon to understand features of the nature of evil that reason alone cannot grasp.From the Preface:'By combining several disciplines—philosophy, anthropology and literary criticism, as well as psychoanalysis—Flahault scrutinizes the origin of malevolence and reveals that, contrary to the view presented by moral philosophy, it is within us that the roots of wickedness are to be found ... Taking issue with the widely accepted view that monotheism constitutes moral progress, he argues that by instigating a dualism between good and evil, monotheism has in fact foreclosed the possibility of acknowledging the ambivalence of our fascination with the limitless and infinity.' Chantal Mouffe.
DPT (Ours)	Nonfiction-Religion & Philosophy-Philosophy
Multi-Verb (Baseline)	Nonfiction-Religion & Philosophy-Philosophy, Nonfiction-Religion & Philosophy- Religion
Text	Garfield Caution: Wide Load: Indulge the Bulge Garfield believes that a full belly is a happy belly—and he intends to keep his stomach ecstatic. Fans of the fat cat will gleefully fill up on this latest smorgasbord of fun!
DPT (Ours)	Humor
Multi-Verb (Baseline)	Humor-Graphic Novels & Manga (Label inconsistency)

Table 10: Case Studies on BGC dataset.