SEP-MLDC: A Simple and Effective Paradigm for Multi-Label Document Classification

Han Liu[1], **Shuqin Li**[1], **Xiaotong Zhang**[1*], **Yuanyuan Wang**[1],
Feng Zhang[2], **Hongyang Chen**[3], **Hong Yu**[1]
[1]Dalian University of Technology, Dalian, China
[2]Peking University, Beijing, China
[3]Zhejiang Lab, Hangzhou, China
{liu.han.dut,zfeng.maria}@gmail.com, {shuqinli,wangy9yuan}@mail.dlut.edu.cn,
zxt.dut@hotmail.com, dr.h.chen@ieee.org, hongyu@dlut.edu.cn

Abstract

Multi-label document classification (MLDC) aims to allocate more than one label to each document and attracts increasing attention in many practical applications. However, previous studies have failed to pay sufficient attention to the lack of semantic information on labels and the long-tail problem prevalent in the datasets. Additionally, most existing methods focus on optimizing document features, overlooking the potential of high-quality label features to enhance classification performance. In this paper, we propose a simple and effective paradigm for MLDC. Regarding the problem of insufficient label information and imbalance in the sample size of categories, we utilize large language models (LLMs) to semantically expand the label content and generate pseudo-samples for the tail categories. To optimize the features of both documents and labels, we design the contrastive learning boosted feature optimization module facilitated by the similarity matrices. Finally, we construct a label-guided feature selection module to incorporate the optimized label features into the input features to provide richer semantic information for the classifier. Extensive experiments have demonstrated that our proposed method significantly outperforms state-of-the-art baselines.

1 Introduction

Multi-label document classification (MLDC) has attracted significant attention and in-depth study due to its wide range of applications across various fields, such as recommendation systems (Zhang et al., 2019), sentiment analysis (Kamila et al., 2022), and user profiling (Wen et al., 2023).

Previous MLDC methods have achieved promising results but still suffer from the following three limitations. First, most of them fail to fully utilize the semantic information of the labels. Although Zhang et al. (2023) attempt to use label names as the inputs of the backbone models to obtain label features, the problem is that label names are usually concise. In some datasets (e.g., EUR-Lex (Mencía and Fürnkranz, 2008)), most labels are represented as IDs, leading to a serious lack of semantic information. Second, current studies require sophisticated designs to address long-tail problems, posing challenges in terms of feasibility and adaptability to diverse scenarios. Hüllermeier et al. (2022) enhance the model performance on tail labels by optimizing the loss function to allocate more attention to the tail categories. Such methods are relatively complex to implement and may not be suitable for all scenarios. Sang et al. (2022) leverage the concept of transfer learning to acquire general knowledge from the head-labels and then transfer this knowledge to the tail-labels with fewer samples. This approach typically requires considerable effort in designing specialized modules for feature transfer. Lastly, most current research concentrates on enhancing the document features while neglecting the fact that label features can also significantly improve classification performance. For example, Xu et al. (2023) decompose multi-label samples into multiple single-label samples, compute label prototypes based on these single-label samples, and optimize their embeddings in the latent space using contrastive learning. Although Bai et al. (2022) try to optimize latent representations through contrastive learning based on the similarity between label features and document features. It ignores quantitative information related to labels and documents, such as label co-occurrence and label consistency across documents.

To address the problems mentioned above, we propose an innovative framework called SEP-MLDC, which optimizes document and label features simultaneously. The framework is shown in Figure 1, which consists of four modules: the label semantic enhancement module, the data augmentation module, the feature optimization module, and

[*]Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3849–3859
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

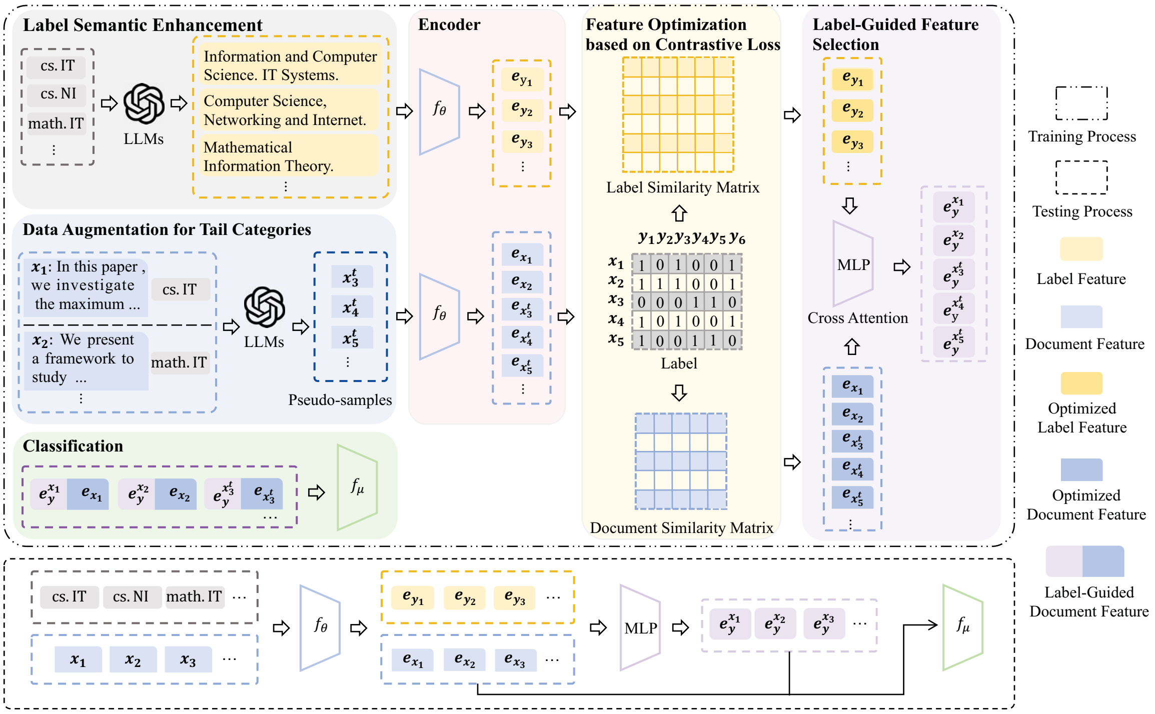
Label Semantic Enhancement
cs. IT
cs. NI
math. IT
LLMs
Information and Computer Science. IT Systems.
Computer Science, Networking and Internet.
Mathematical Information Theory.
Data Augmentation for Tail Categories
x_1: In this paper, we investigate the maximum ...
cs. IT
x_2: We present a framework to study ...
math. IT
LLMs
x_3^t
x_4^t
x_5^t
Pseudo-samples
Classification
$e_y^{x_1}$ e_{x_1} $e_y^{x_2}$ e_{x_2} $e_y^{x_3}$ e_{x_3}
f_μ
Encoder
f_θ
e_{y_1} e_{y_2} e_{y_3}
e_{x_1} e_{x_2} $e_{x_3^t}$ $e_{x_4^t}$ $e_{x_5^t}$
Feature Optimization based on Contrastive Loss
Label Similarity Matrix
$y_1\ y_2\ y_3\ y_4\ y_5\ y_6$
x_1 1 0 1 0 0 1
x_2 1 1 1 0 0 1
x_3 0 0 0 1 1 0
x_4 1 0 1 0 0 1
x_5 1 0 0 1 1 0
Label
Document Similarity Matrix
Label-Guided Feature Selection
MLP
Cross Attention
$e_y^{x_1}$ $e_y^{x_2}$ $e_y^{x_3^t}$ $e_y^{x_4^t}$ $e_y^{x_5^t}$
Training Process
Testing Process
Label Feature
Document Feature
Optimized Label Feature
Optimized Document Feature
Label-Guided Document Feature
cs. IT cs. NI math. IT ...
x_1 x_2 x_3 ...
f_θ
e_{y_1} e_{y_2} e_{y_3} ...
e_{x_1} e_{x_2} e_{x_3} ...
MLP
$e_y^{x_1}$ $e_y^{x_2}$ $e_y^{x_3}$ ...
f_μ

Figure 1: The architecture of the proposed framework. The arrows in the figure show the data flow in the model.

the label-guided feature selection module. First, the label semantic enhancement module aims to solve the problem of insufficient label semantic information. In many MLDC datasets, labels may appear as IDs or abbreviations, e.g., "7937.0" in the EUR-Lex dataset and "cs.IR" in the AAPD dataset (Yang et al., 2018). We leverage the rich knowledge of LLMs to provide semantic interpretations for labels, thereby obtaining more detailed content for labels. Second, we employ the data augmentation module to tackle the long-tail problem. Specifically, we leverage existing LLMs to generate pseudo-samples for tail categories, enhancing data in these categories and alleviating the effects of severe class imbalance. Third, we devise a contrastive learning enhanced feature optimization module based on two similarity matrices to obtain superior document features and label features. We construct a document similarity matrix derived from label relationships and a label similarity matrix derived from document relationships to optimize document features and label features based on contrastive learning respectively.

Finally, we aim to integrate label features with document features to enhance the classifier's expressive power. We design a label-guided feature selection module that generates a specific representation based on each label's importance to the document for the final classification.

Our contributions are as follows: (1) We leverage LLMs to enhance label semantic information and generate pseudo-samples for the tail categories. (2) We propose a novel similarity matrices-enhanced contrastive learning approach to optimize document features and label features that are more discriminative for classification. (3) We design a label-guided feature selection module that uses optimized label features and document features to generate representations for the final classification, enriching the classifier's input features. (4) Experimental results show that our model can achieve state-of-the-art performance in comparison with other strong baselines.

2 Related Work

2.1 Multi-Label Document Classification

Multi-label classification (MLC) methods primarily focus on enhancing representation learning (Liu et al., 2017) and modeling label dependencies (Yang et al., 2018; Tsai and Lee, 2020; Fallah et al., 2023; Du et al., 2024). For instance, Du et al. (2024) propose LD-SPN, a multi-label text classification method based on Set Prediction Networks (SPN). It utilizes Graph Convolutional Networks (GCN) to model label dependencies and introduces the Bhattacharyya Distance to optimize the output distributions. However, it neglects the fact that different labels may concentrate on distinct tokens. Therefore, label-specific feature learning (You et al., 2019; Xiao et al., 2019; Ma et al., 2021; Zhang et al., 2021b,a; Mao et al., 2023), which captures the unique characteristics of each label, is a promising approach for improving label differentiation. You et al. (2019) propose a label-specific attention network that attends to different tokens when predicting each label. But it neglects the exploration of label relationships. Additionally, Ma et al. (2021) utilize GCN to integrate label information and model-adaptive interactions in a label-specific manner. Zhang et al. (2021a) leverage correlation-guided representations to capture high-order document-label correlations. Bai et al. (2022) employ Variational Autoencoder (VAE) to learn and align the embedding spaces of labels and documents. However, these methods overlook the significant impact of label semantic information on the model and fail to leverage the potential benefits of jointly optimizing both document and label features.

2.2 Contrastive Learning

Contrastive learning (Oord et al., 2018; Chen et al., 2020; Khosla et al., 2020) is an effective approach for optimizing the feature space. The core idea is to make an anchor sample close to its similar samples (positive samples) and far from dissimilar samples (negative samples) in the embedding space. Contrastive loss is largely inspired by Noise Contrast Estimation (NCE) (Gutmann and Hyvärinen, 2010) and its form is generalizable. The original contrastive loss only considers instance-level invariance, using multiple views of the anchor instance as its positive set. In this paper, we leverage label information to construct a similarity matrix for documents enabling supervised contrastive learning on document features. Conversely, document information is used to construct a similarity matrix for labels, facilitating supervised contrastive learning on label features. Document and label features are optimized simultaneously to provide the classification model with higher-quality feature inputs.

3 The Proposed Method

3.1 Problem Definition

We use $\mathcal{D}$ to denote the training set, prompt_l and $\mathrm{prompt}_{\mathcal{D}}$ represent the prompts for expanding the semantic information of the labels and to generate the pseudo-samples for tail categories, respectively. Individual documents in the dataset are denoted by $\boldsymbol{x}$. $\mathcal{Y}$ represents the label space associated with the dataset. For each document $\boldsymbol{x}_i$, it corresponds to a set of labels $\boldsymbol{y}_i = [y_1, \ldots, y_j, \ldots, y_L]$, where $y_j \in \{0, 1\}$, and $y_j = 1$ indicates that the j-th label is correlated with the document $\boldsymbol{x}_i$. L represents the total number of all candidate labels. In the testing phase, our core task is to predict all possible relevant labels for a new document.

3.2 Label Semantic Enhancement

Some labels in the dataset are represented as IDs or acronyms with ambiguous meanings, significantly hindering the model's ability to learn label semantic information effectively. To enrich label information, we design efficient prompts to guide the LLM in expanding label content:

$$y_j^* = \mathrm{LLM}(\mathrm{prompt}_l, y_j),\ y_j \in \mathcal{Y}, \quad (1)$$

where the LLM used here is GPT-4. We then combine the original label with its expanded content to create a more semantically explicit expression, which will be fed into the backbone model. This enables the model to capture label semantics more comprehensively and accurately:

$$y_j' = \left[y_j; y_j^*\right], y_j \in \mathcal{Y}. \quad (2)$$

The expanded labels now contain richer semantic content, allowing the model to better understand their inherent meaning. Some examples of expanded labels used in our dataset are provided in Appendix A.

3.3 Data Augmentation for Tail Categories

The imbalanced class distribution in multi-label datasets causes a severe long-tail phenomenon. This makes models more likely to focus on head categories with larger sample sizes while relatively neglecting tail categories with fewer samples.

To enhance the model's ability to learn the tail category information more effectively, we leverage existing LLMs to generate pseudo-samples for tail categories $\mathcal{Y}_{\text{tail}}$, defined as categories with fewer than 50 samples. Specifically, we use dataset information and expanded label content as prompts, along with a sample from the corresponding category as a demonstration. We then generate 50 pseudo-samples for each tail category and incorporate them into the training set for subsequent model training:

$$\boldsymbol{x}_{y_j}^t = \mathrm{LLM}(prompt_{\mathcal{D}}, y_j, \boldsymbol{x}_{y_j}),\ y_j \in \mathcal{Y}_{\text{tail}}, \quad (3)$$

$$\mathcal{D}_{\text{tail}} = \{\boldsymbol{x}_{y_j}^t\},\ y_j \in \mathcal{Y}_{\text{tail}}, \tag{4}$$

where $\boldsymbol{x}_{y_j}$ denotes the sample of the tail category y_j, $\boldsymbol{x}_{y_j}^t$ represents the pseudo-sample of the tail category t generated by the LLM. Here, we utilize GPT-4 as the LLM. In $\text{prompt}_{\mathcal{D}}$, we provide dataset information and specify the requirements for the expected pseudo-samples. The specific prompts used in our study are provided in Appendix B.

3.4 Similarity Matrices Enhanced Contrastive Learning

Label Similarity Matrix Based on Document Consistency Refining label features in the latent space by increasing the distance between irrelevant labels and keeping relevant labels closer benefits model classification.

We utilize supervised contrastive learning to optimize the label representations in the latent space. A label similarity matrix is constructed based on the relationship between different labels that represent the same document and is applied to contrastive learning to refine label features.

In contrastive learning, constructing the positive set and negative set for the anchor document is crucial. We treat the anchor label y_i as the positive sample. Other labels $\{y_j\}, y_j \in \mathcal{N}_y$ are considered as negative samples for label y_i. Here, $\mathcal{N}_y$ denotes the negative set for label y_i.

The label similarity matrix $\mathbf{W}_y \in \mathbb{R}^{bs\times bs}$ is defined as:

$$\mathbf{W}_y = \mathbf{L}_{same} \circ \mathbf{A}_y, \tag{5}$$

$$\mathbf{A}_{y_i} = \mathbf{S}_{y_i}\mathbf{E} + \mathbf{S}_y - \mathbf{L}_{same_i}, \tag{6}$$

$$\mathbf{L}_{same} = \mathbf{Y}^{\mathbf{T}}\mathbf{Y}, \tag{7}$$

$$\mathbf{S}_y = \text{sum}(\mathbf{Y}). \tag{8}$$

$\mathbf{Y} \in \mathbb{R}^{bs\times L}$ represents the label matrix of the documents, where bs is the batch size and L denotes the number of labels. The operator sum$(\cdot)$ refers to summing matrix rows.

$\mathbf{S}_y \in \mathbb{R}^{bs}$ denotes the number of labels assigned to each document. The element $\mathbf{L}_{same_{ij}}$ in $\mathbf{L}_{same} \in \mathbb{R}^{bs\times bs}$ indicates the number of shared labels between document $\boldsymbol{x}_i$ and document $\boldsymbol{x}_j$. The vector $\mathbf{L}_{same_i}$ corresponds to the i-th row of the matrix $\mathbf{L}_{same}$.

Additionally, $\mathbf{E}$ denotes the identity matrix. The element $\mathbf{A}_{y_{ij}}$ in $\mathbf{A}_{y_i} \in \mathbb{R}^{bs}$ represents the number of documents assigned both label y_i and label y_j. The operator $\circ$ denotes the element-wise division between two matrices.

The contrastive learning loss of labels is defined as $\mathcal{L}_1$:

$$\mathcal{L}_1=\frac{1}{L}\sum_{y_i\in\mathcal{Y}}-\log\left(\frac{\exp(\boldsymbol{e}_{y_i}\cdot\boldsymbol{e}_{y_i}/\tau)\mathbf{W}_{y_{ii}}}{\sum\limits_{y_k\in\mathcal{N}_y}\exp(\boldsymbol{e}_{y_i}\cdot\boldsymbol{e}_{y_k}/\tau)\mathbf{W}_{y_{ik}}}\right), \tag{9}$$

where τ is the temperature coefficient in contrastive learning, and $\boldsymbol{e}_{y_i} = f_\theta(y_i')$ denotes the embedding of the i-th label. The f_θ used here is the pre-trained Roberta-base model(Liu et al., 2019). The operator $\cdot$ is the dot product between embeddings. The element $\mathbf{W}_{y_{ii}}$ refers to the value at the i-th row and i-th column of the label similarity matrix.

This process encourages a more structured organization of label features in the latent space, ensuring that highly relevant labels remain close while irrelevant labels are distinctly separated.

Document Similarity Matrix Based on Label Coexistence In multi-label document classification, latent space feature representations often become ambiguous and entangled, as these documents correspond to multiple labels simultaneously. This distinction is the fundamental difference between multi-label and single-label classification.

Previous methods employ attention mechanisms to compute label-specific document representations and train models in single-label format. However, these approaches overlook the advantages of incorporating label-relevant information.

In previous methods using contrastive learning, documents sharing common labels, even if it's just one, are typically assigned to the positive set, while all others are classified as negative. However, due to the large number of labels, this kind of division method hinders the learning of high-quality document features, as the inconsistency of non-common labels confuses the model. To address this problem, we design a new strategy where only documents with identical labels are added to the positive set $\mathcal{P}_{\boldsymbol{x}}$, while all other documents are placed in the negative set $\mathcal{N}_{\boldsymbol{x}}$. Specifically, the label set $\boldsymbol{y}_j$ of the document $\boldsymbol{x}_j$ in the positive set of the anchor document $\boldsymbol{x}_i$ must satisfy $\boldsymbol{y}_j = \boldsymbol{y}_i$. For documents in the negative set, we compute a similarity matrix based on the number of labels they share with the anchor document and incorporate it into the contrastive learning process. The document similarity matrix $\mathbf{W}_{\boldsymbol{x}} \in \mathbb{R}^{L\times L}$ is given by:

$$\mathbf{W}_{\boldsymbol{x}} = \mathbf{Doc}_{same} \circ \mathbf{A}_{\boldsymbol{x}}, \tag{10}$$

Dataset	N_{trn}	N_{tst}	D_{vocab}	L	L_{avg}	W_{trn}	W_{tst}
AAPD	54,840	1,000	69,399	54	2.41	163.42	171.65
RCV1	23,149	781,265	47,236	103	3.18	259.47	269.23
EUR-Lex	11,585	3,865	171,120	3,956	5.32	1225.20	1248.07

Table 1: Data statistics. N_{trn}, N_{tst} refer to the number of documents in the training and test sets, respectively. D_{vocab} is the vocabulary size of documents. L is the number of labels. L_{avg} is the average number of labels per document. W_{trn}, W_{tst} refer to the average number of words per document in the training and test sets, respectively.

$$\mathbf{A}_{\boldsymbol{x}_i} = \mathbf{S}_{\boldsymbol{x}_i}\mathbf{E} + \mathbf{S}_{\boldsymbol{x}} - \mathbf{Doc}_{same_i}, \tag{11}$$

$$\mathbf{Doc}_{same} = \mathbf{Y}\mathbf{Y}^{\mathbf{T}}, \tag{12}$$

$$\mathbf{S}_{\boldsymbol{x}} = \text{sum}(\mathbf{Y}^{\mathbf{T}}). \tag{13}$$

The element in $\mathbf{S}_{\boldsymbol{x}} \in \mathbb{R}^L$ represents the number of documents tagged with each label. The element $\mathbf{Doc}_{same_{ij}}$ in $\mathbf{Doc}_{same} \in \mathbb{R}^{L\times L}$ denotes the number of documents tagged by both label y_i and the label y_j. $\mathbf{Doc}_{same_i}$ refers to the i-th row of the matrix $\mathbf{Doc}_{same}$. $\mathbf{E}$ is the identity matrix. The element $\mathbf{A}_{\boldsymbol{x}_{ij}}$ in $\mathbf{A}_{\boldsymbol{x}_i} \in \mathbb{R}^L$ represents the total number of labels after de-emphasis corresponding to documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Similar to the label matrix, we construct the document similarity matrix which is used in the supervised contrastive loss to optimize document features:

$$\mathcal{L}_{x_i} = \sum_{\boldsymbol{x}_j \in \mathcal{P}_{\boldsymbol{x}}} -\log\left(\frac{\exp(\boldsymbol{e}_{\boldsymbol{x}_i}\cdot\boldsymbol{e}_{\boldsymbol{x}_j}/\tau)\mathbf{W}_{\boldsymbol{x}_{ij}}}{\sum\limits_{\boldsymbol{x}_k \in \mathcal{N}_{\boldsymbol{x}}}\exp(\boldsymbol{e}_{\boldsymbol{x}_i}\cdot\boldsymbol{e}_{\boldsymbol{x}_k}/\tau)\mathbf{W}_{\boldsymbol{x}_{ik}}}\right), \tag{14}$$

where $\boldsymbol{e}_{\boldsymbol{x}_i} = f_\theta(\boldsymbol{x}_i)$ represents the embedding of the i-th document, and $\mathbf{W}_{\boldsymbol{x}_{ij}}$ is the element in the i-th row and j-th column of the document similarity matrix. Given a batch of documents $\mathcal{B}$, the contrastive loss of documents can be expressed as:

$$\mathcal{L}_2 = \frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{x}_i \in \mathcal{B}} \mathcal{L}_{x_i}. \tag{15}$$

Different representation distances are established based on the ratio of shared labels among documents, effectively alleviating the issue of document representations being easily confused in the feature space.

By constructing the two similarity matrices, we guide the model to optimize both label and document features in the latent space through contrastive learning.

3.5 Label-Guided Feature Selection

In previous studies, researchers have concatenated label features with document features and input them into the model. This was done in an attempt to enhance the predictive power of the classification model through a more comprehensive representation. However, it has a significant limitation during the testing phase: the absence of labels forces researchers (Zhang et al., 2023) to replace the label part with $\mathbf{0}$, which inevitably weakens the interpretability of the model. Moreover, in multi-label classification, directly concatenating labels with document data is impractical due to the varying number of labels associated with each document.

Inevitably, redundant information that is not strongly correlated with the labels appears in the document. To address this, we employ a label-guided content querying mechanism to extract high-quality features from the document features $\boldsymbol{e}_{\boldsymbol{x}} \in \mathbb{R}^{bs\times d}$ and label features $\boldsymbol{e}_{\boldsymbol{y}} \in \mathbb{R}^{L\times d}$ which are optimized in Section 3.4.

We hope to learn a weight to reflect the importance of different labels to the document and use it to aggregate the label features to obtain a document-specific label representation:

$$\boldsymbol{e}_y^{\boldsymbol{x}} = \alpha\boldsymbol{e}_y, \tag{16}$$

$$\alpha = \text{softmax}(\boldsymbol{e}_{\boldsymbol{x}}\boldsymbol{e}_y^{\mathbf{T}}). \tag{17}$$

Next, we concatenate $\boldsymbol{e}_y^{\boldsymbol{x}}$ with the document feature to create the input for the classifier f_μ, which is a multi-layer perception:

$$\boldsymbol{e}_{input} = [\boldsymbol{e}_y^{\boldsymbol{x}}; \boldsymbol{e}_{\boldsymbol{x}}]. \tag{18}$$

Finally, we apply cross-entropy loss to all the documents. In the multi-label setting, since a document can have multiple labels, we must consider all candidate labels for each document $\boldsymbol{x}_i$. The classification loss is calculated as:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathbb{D}|}\sum_{\boldsymbol{x}_i \in \mathbb{D}}\sum_{j=1}^{L} y_j \log p(y_{pre}=y_j|\boldsymbol{e}_{input}), \tag{19}$$

Method	P@1	P@3	P@5	N@3	N@5
XML-CNN	74.38	53.84	37.79	71.12	75.93
SGM	75.67	56.75	35.65	72.36	75.30
DXML	80.54	56.30	39.16	77.23	80.99
Attn-XML	83.02	58.72	40.56	78.01	82.31
EXAM	83.26	59.77	40.66	79.10	82.79
LSAN	85.28	61.12	41.84	80.84	84.78
HTTN	83.84	59.92	40.79	79.27	82.67
LDGN	86.24	61.95	42.29	83.32	86.85
LSFA	86.95	62.88	**43.43**	83.96	87.53
SEP-MLDC	**90.30**	**64.56**	43.08	**85.72**	**88.93**

Table 2: Comparison results on the AAPD dataset.

Method	P@1	P@3	P@5	N@3	N@5
XML-CNN	95.75	78.63	54.94	89.89	90.77
SGM	95.37	81.36	53.06	91.76	90.69
DXML	94.04	78.65	54.38	89.83	90.21
Attn-XML	96.41	80.91	56.38	91.88	92.70
EXAM	93.67	75.80	52.73	86.85	87.71
LSAN	96.81	81.89	56.92	92.83	93.43
HTTN	95.86	78.92	55.27	89.61	90.86
LDGN	97.12	82.26	57.29	93.80	95.03
LSFA	97.21	82.52	57.52	94.20	**95.42**
SEP-MLDC	**98.06**	**85.64**	**61.36**	**94.33**	94.92

Table 3: Comparison results on the RCV1 dataset.

Method	P@1	P@3	P@5	N@3	N@5
XML-CNN	70.40	54.98	44.86	58.62	53.10
SGM	70.45	60.37	43.88	60.72	55.24
DXML	75.63	60.13	48.65	63.96	53.60
Attn-XML	79.66	64.88	52.99	68.66	62.33
EXAM	74.40	61.93	50.98	65.12	59.43
LSAN	79.17	64.99	53.67	68.32	62.47
HTTN	80.45	65.57	55.68	69.01	63.76
LDGN	81.03	67.79	56.36	71.81	66.09
LSFA	83.75	70.74	58.95	74.13	68.25
SEP-MLDC	**85.88**	**71.94**	**59.52**	**76.16**	**68.83**

Table 4: Comparison results on the EUR-Lex dataset.

$$\mathbb{D} = \mathcal{D} \cup \mathcal{D}_{\text{tail}}, \tag{20}$$

where y_{pre} is the label predicted by the model. $\mathbb{D}$ denotes the set of original documents $\mathcal{D}$ and documents generated for the tail categories $\mathcal{D}_{\text{tail}}$. By combining equations (9), (15) and (19), the overall loss of our proposed framework is:

$$\mathcal{L} = \mathcal{L}_{ce} + \beta\mathcal{L}_1 + \gamma\mathcal{L}_2, \tag{21}$$

where β and γ are adjustable weight parameters.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics We evaluate the proposed model on three MLTC datasets including AAPD (Yang et al., 2018), RCV1 (Lewis et al., 2004), and EUR-Lex (Mencía and Fürnkranz, 2008). Table 1 contains the statistics of these three benchmark datasets. We follow the settings of previous works (You et al., 2019; Ma et al., 2021; Xiao et al., 2021; Xu et al., 2023) to use precision at $k(P@k)$ and normalized discounted cumulative gain at $k(N@k)$ as evaluation metrics.

Parameter Settings We set the loss function parameters β and γ to 0.01. The temperature coefficient τ in contrastive learning is set to 0.1. The AdamW optimizer (Loshchilov and Hutter, 2019) is used with an initial learning rate of 5×10^{-5} and a dropout rate of 0.5. All experiments are conducted on a single 3090 GPU, and the experimental results for each dataset are obtained by averaging the results of ten trials.

4.2 Baselines

We compare our method with the following strong baselines. **XML-CNN** (Liu et al., 2017) is a CNN-based model using a dynamic max pooling scheme to capture high-level features. **SGM** (Yang et al., 2018) is a sequence generation model that models the correlations between labels. **DXML** (Zhang et al., 2018) is a deep embedding method that models the feature and label space simultaneously. **Attn-XML** (You et al., 2019) is a deep learning model that uses multi-label attention to extract information for each label. **EXAM** (Du et al., 2019) is a framework that employs the interaction mechanism to compute the word-level interaction signals. **LSAN** (Xiao et al., 2019) is a label-specific attention model based on self-attention and label-attention mechanisms. **HTTN** (Xiao et al., 2021) is a head-to-tail network that transfers the meta-knowledge from the head-labels to the tail-labels. **LDGN** (Ma et al., 2021) is a graph convolution network that incorporates category information and models adaptive interactions of labels. **LSFA** (Xu et al., 2023) is a prototype-based VAE-style feature generation model to capture the intra-class semantic variations from the head-labels and then apply it to augment features for tail-labels.

4.3 Result Analysis

Tables 2, 3 and 4 report the experimental results for the MLDC tasks on AAPD, RCV1, and EUR-Lex datasets respectively. The best results are highlighted in bold. From the results, we observe that SEP-MLDC performs significantly bet-

Dataset	AAPD					EUR-Lex				
Modules	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
SEP-MLDC w/o LSE	85.2	61.29	39.43	82.02	85.66	76.44	65.73	52.94	71.24	63.23
SEP-MLDC w/o DATC	88.61	61.87	43.16	84.02	86.10	75.19	64.27	50.55	68.39	56.12
SEP-MLDC w/o SMCL	82.12	59.27	41.37	80.29	83.28	78.97	68.14	55.76	72.48	62.60
SEP-MLDC w/o LGFS	84.73	60.19	40.16	82.84	85.72	80.31	69.46	57.58	72.19	65.79
SEP-MLDC	**90.30**	**64.56**	43.08	**85.72**	**88.93**	**85.88**	**71.94**	**59.52**	**76.16**	**68.83**

Table 5: Ablation study on key components in SEP-MLDC. LSE denotes the label semantic enhancement module, DATC stands for the module of data augmentation for tail categories, SMCL represents the feature optimization module based on similarity matrices-enhanced contrastive learning, and LGFS refers to the label-guided feature selection module.

ter than the other baselines, demonstrating the superiority of our method. Specifically, for the AAPD dataset (Table 2), SEP-MLDC improves upon the most competitive baseline, LSFA, by 3.35%, 1.68%, 1.76%, and 1.40% in terms of P@1, P@3, N@3, N@5 scores, respectively. For the RCV1 dataset (Table 3), SEP-MLDC improves upon LSFA by 0.85%, 3.12%, 3.84%, and 0.13% in terms of P@1, P@3, P@5, N@3 scores. And for the EUR-Lex dataset (Table 4), SEP-MLDC outperforms LSFA by 2.13%, 1.20%, 0.57%, 2.03%, and 0.58% on all metrics.

The main reason SEP-MLDC achieves such a performance improvement is that it leverages LLMs to semantically enrich the labels. This process gives the labels a clearer and richer meaning within the dataset, greatly enhancing their expressive power. Additionally, LLMs are utilized to generate pseudo-samples for tail categories. This encourages the model to focus more on categories with fewer samples, thereby mitigating the impact of the long-tail problem. Furthermore, effectively modeling both label relevance and document relevance allows the model to learn more discriminative feature representations. Lastly, SEP-MLDC integrates the label information that documents emphasize into the features to be classified. This innovative approach significantly enhances the information density of the features, further improving the model's performance.

4.4 Ablation Study

We analyze the impacts of key components in SEP-MLDC. Table 5 shows the results evaluated on AAPD and EUR-Lex datasets.

Label Semantic Enhancement After removing the label semantic enhancement module, we directly input the original labels into the encoder to obtain label features for subsequent modules. As shown in Table 5, it is clear that the absence of label semantic information leads to a significant drop in model performance on both the AAPD and EUR-Lex datasets. This is especially evident in the EUR-Lex dataset, where the absence of label semantic information notably hinders performance due to the large number of ambiguous labels, highlighting the importance of label semantic enhancement.

Data Augmentation for Tail Categories As shown in Table 5, the absence of the data augmentation module for tail categories leads to a noticeable decline in model performance on both the EUR-Lex and AAPD datasets. Notably, this effect is more pronounced in the EUR-Lex dataset, where P@1 drops by as much as 10.69%. This can be attributed to the more severe long-tail problem in EUR-Lex compared to AAPD, as the EUR-Lex dataset contains a significantly larger number of tail categories. From this perspective, the tail category data augmentation module plays a crucial role in effectively mitigating the long-tail issue.

Similarity Matrices Enhanced Contrastive Learning Table 5 shows a notable decrease in model performance when the two similarity matrices module is removed, with a drop of 8.18% in the P@1 metric for the AAPD dataset and 6.91% for the EUR-Lex dataset, underscoring its importance. By modeling both label and document relationships, we construct a label similarity matrix and a document similarity matrix. These matrices are used in the corresponding feature contrastive learning to optimize label and document representations, thereby making the feature distribution in the latent space more coherent.

Label-Guided Feature Selection From Table 5, we observe that the label-guided feature selection module is second only to the similarity matrices-

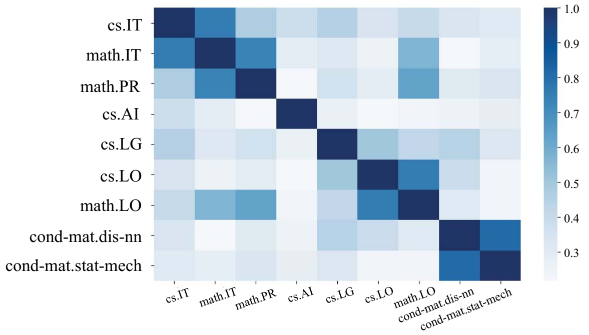

Figure 2: Label-label inner-products.

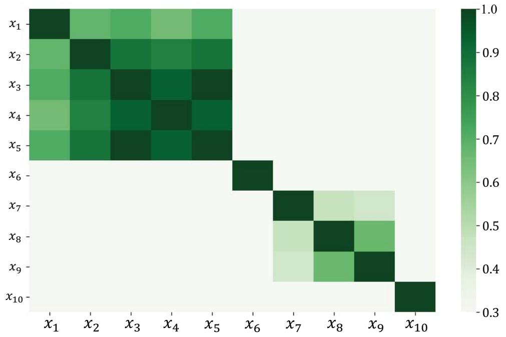

Figure 3: Document-document inner-products.

enhanced feature optimization module in terms of importance. This module enriches the input features by providing more valuable information, thereby improving classification performance.

4.5 Performance on Tail Categories

To further evaluate the effectiveness of the SEP-MLDC in alleviating the long-tail problem, we compare its performance against other baselines using PSP@k, which amplifies the impact of rare labels by applying inverse propensity weighting. It adjusts the emphasis on rare labels by assigning them higher weights.

We conduct performance experiments on tail categories in the EUR-Lex dataset, where the long-tail issue is particularly severe. Table 6 shows that SEP-MLDC significantly outperforms the baselines in tail label classification. This is expected, as SEP-MLDC effectively addresses the data sparsity of tail labels by leveraging LLMs to generate pseudo-samples for these labels.

Method	PSP@1	PSP@3	PSP@5
LSAN	36.41	41.27	43.42
HTTN	38.96	43.28	45.74
LSFA	42.50	48.03	50.69
SEP-MLDC	**44.93**	**51.84**	**53.78**

Table 6: Performance on tail-categories on the EUR-Lex dataset.

4.6 Visualization

We conduct a thorough analysis of the label representations learned by the model, as shown in Figure 2. This figure illustrates a map of the inner-product weights of partial label embeddings from the AAPD dataset. The heatmap visualization provides an intuitive representation of the relationships among labels, where the intensity of the colors directly reflects the strength of their correlation. Notably, labels like "cs.IT" and "math.IT" are often annotated together on the same documents. These labels are represented by darker shades on the map, indicating a high degree of co-occurrence and strong correlation. In contrast, labels like "math.PR" and "cs.AI", which never appear together in the training set, exhibit low feature similarity and are depicted in lighter hues. This observation suggests that the learned label embeddings have effectively captured the inter-label correlations. Additionally, the optimized label features serve as a beneficial adjunct for selecting document features.

As shown in Figure 3, a heatmap depicts the feature similarity among samples from the AAPD dataset. The heatmap reveals that the first five documents exhibit a high degree of feature similarity, whereas $\boldsymbol{x}_7$, $\boldsymbol{x}_8$, and $\boldsymbol{x}_9$ from another cluster characterized with elevated similarity. In contrast, $\boldsymbol{x}_6$ and $\boldsymbol{x}_{10}$ are observed in isolation, exhibiting minimal feature similarity and little correlation with the other documents. Notably, the label sets of $\boldsymbol{x}_6$ and $\boldsymbol{x}_{10}$ are entirely distinct from the others, with no overlapping labels. It is worth noting that when the feature similarity between the documents exceeds or equals a threshold value of 0.8, their label sets are found to be identical. These observations confirm that the discriminability of document features within the latent space is significantly enhanced after the feature optimization. The problem of feature ambiguity is alleviated. Specifically, the similarity matrices-enhanced contrastive learning module helps bring document features with similar label sets closer together.

5 Conclusion

In this paper, we first introduce a semantic expansion strategy for labels through LLMs, aiming to enrich their content and diversity. With the assistance of LLMs, we enhance the number of tail categories by generating pseudo-samples for the tail categories. Then we propose the similarity matrices, an innovative approach applied in contrastive learning, to optimize the feature representation of both documents and labels. The label-guided feature selection module is designed to fuse relevant label features for the document, enriching the information and aiding model classification. Experiments demonstrate that our method outperforms state-of-the-art approaches.

Limitation

In our approach, we utilize LLMs to semantically enrich labels and generate samples for tail categories, which inevitably increases the experimental cost in terms of both time and financial investment. Moreover, the design of the prompt has a significant impact on the quality of the tail category samples generated by LLMs. In future work, we plan to explore the use of open-source large language models. Our goal is to strike a balance between time and computational costs by employing techniques such as batch processing and fine-tuning. Additionally, we will develop a set of prompt templates suitable to various multi-label datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62206038, 62106035), Liaoning Binhai Laboratory Project (No. LBLF-2023-01), Chunhui Project Foundation of the Education Department of China (No. HZKY20220419), and Xiaomi Young Talents Program. We also would like to thank Dalian Ascend AI Computing Center and Dalian Ascend AI Ecosystem Innovation Center for providing inclusive computing power and technical support.

References

Junwen Bai, Shufeng Kong, and Carla P. Gomes. 2022. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning (ICML)*, pages 1383–1398.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607.

Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6359–6366.

Xinkai Du, Quanjie Han, Yalin Sun, Chao Lv, and Maosong Sun. 2024. Label dependencies-aware set prediction networks for multi-label text classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 11206–11210.

Haytame Fallah, Emmanuel Bruno, Patrice Bellot, and Elisabeth Murisasco. 2023. Exploiting label dependencies for multi-label document classification using transformers. In *ACM Symposium on Document Engineering (DocEng)*, pages 12:1–12:4.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 297–304.

Eyke Hüllermeier, Marcel Wever, Eneldo Loza Mencía, Johannes Fürnkranz, and Michael Rapp. 2022. A flexible class of dependence-aware multi-label loss functions. *Mach. Learn.*, 111(2):713–737.

Sabyasachi Kamila, Walid Magdy, Sourav Dutta, and MingXue Wang. 2022. AX-MABSA: A framework for extremely weakly supervised multi-label aspect based sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6136–6147.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 18661–18673.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 115–124.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 3855–3864.

Junxiang Mao, Wei Wang, and Min-Ling Zhang. 2023. Label specific multi-semantics metric learning for multi-label classification: Global consideration helps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4055–4063.

Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. *Springer-Verlag*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jianghui Sang, Yongli Wang, Long Yuan, Hao Li, and Xiaohui Jiang. 2022. Multi-label transfer learning via latent graph alignment. *World Wide Web*, 25(2):879–898.

Che-Ping Tsai and Hung-yi Lee. 2020. Order-free learning alleviating exposure bias in multi-label classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6038–6045.

Xiang Wen, Shiwei Zhao, Haobo Wang, Runze Wu, Manhu Qu, Tianlei Hu, Gang Chen, Jianrong Tao, and Changjie Fan. 2023. Multi-source multi-label learning for user profiling in online games. *IEEE Transactions on Multimedia (TMM)*, 25:4135–4147.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 466–475.

Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does head label help for long-tailed multi-label text classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 14103–14111.

Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. 2023. Label-specific feature augmentation for long-tailed multi-label text classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 10602–10610.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *International Conference on Computational Linguistics (COLING)*, pages 3915–3926.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5812–5822.

Feng Zhang, Wei Chen, Fei Ding, and Tengjiao Wang. 2023. Dual class knowledge propagation network for multi-label few-shot intent detection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8605–8618.

Qian-Wen Zhang, Ximing Zhang, Zhao Yan, Ruifang Liu, Yunbo Cao, and Min-Ling Zhang. 2021a. Correlation-guided representation for multi-label text classification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3363–3369.

Suwei Zhang, Yuan Yao, Feng Xu, Hanghang Tong, Xiaohui Yan, and Jian Lu. 2019. Hashtag recommendation for photo sharing services. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 5805–5812.

Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha. 2018. Deep extreme multi-label learning. In *International Conference on Multimedia Retrieval (ICMR)*, pages 100–107.

Yanyi Zhang, Xinyu Li, and Ivan Marsic. 2021b. Multi-label activity recognition using activity-specific features and activity correlations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14625–14635.

A Examples of Label Semantic Enhancement

Table 7 shows the prompts we constructed for semantic expansion of the AAPD and EUR-Lex labels, along with some of the expanded label content.

B Prompts of Data Augmentation for Tail Categories

In Table 8, we demonstrate the data for tail labels generated with the assistance of LLMs, as well as the prompts employed.

Dataset	Prompt_Label	Label_Original	Label_Improved
AAPD	The AAPD dataset is a typical multi-label text classification dataset sourced from ArXiv academic papers, containing titles and abstracts, with ArXiv's subject categories as labels. Please explain the meaning of $\{label\}$ in the AAPD dataset in one sentence, using no more than 30 tokens and unrelated information.	stat.ML	The content of the paper is related to the fields of statistical machine learning.
		math.PR	Probability in mathematics, focusing on the study of probability theory and stochastic processes.
		cs.CE	The field of probability in mathematics, focusing on the study of probability theory and stochastic processes.
		physics.data-an	The content of the paper involves data analysis, statistical methods, and probability in the field of physics, encompassing the study and interpretation of physical phenomena through statistical techniques or data-driven approaches.
		quant-ph	Related to the field of quantum physics, covering various directions such as theoretical research, experimental techniques, and quantum information science.
EUR-Lex	The EUR-Lex dataset is a legal document dataset used for multi-label text classification, containing legal texts, regulations, and other legal documents from the European Union. Each document in the dataset is associated with one or more labels that represent different areas of law. Please explain the meaning of $\{label\}$ in the EUR-Lex dataset in one sentence, using no more than 30 tokens and unrelated information.	7937.0	The legal topic "agricultural structure" under the category of "Agriculture and Fisheries".
		accession_criteria	The specific conditions or standards that a country must meet to join the European Union.
		access_to_the_courts	The legal topic concerning the right or ability of individuals or entities to seek judicial redress or remedy in courts.
		accounting_system	Laws and regulations related to the system of recording and summarizing business and financial transactions.
		administrative_expenditure	Legal topics related to the costs of running a government or organization, including salaries, supplies, and overhead expenses.

Table 7: Examples of label semantic enhancement. In the prompt, the placeholder $\{label\}$ should be replaced with the corresponding label from the dataset when it is used.

Dataset	Prompt_Data
AAPD	Example: $\{document\}$. Please refer to the above example and then write an article abstract meeting the following criteria: 1. The abstract belongs to the fields of $\{label_1\}, \ldots, \{label_n\}$. $\{label_1\}$ represents $\{text_1\}$, $\ldots$, $\{label_n\}$ represents $\{text_n\}$. 2. The length is approximately $\{m\}$ tokens. 3. Keep a similar essay writing style, requiring rigor and no mistakes. 4. Describe from several aspects: research background, problems to be solved, proposed methods, and results.
EUR-Lex	Example: $\{document\}$. Please refer to the above example and then write an EU legal document meeting the following criteria: 1. Documents usually include legal provisions, regulations, directives, resolutions, rulings, etc. Each document contains the content of the legal text (such as title, summary or complete legal text). 2. The legal document belongs to the $\{label_1\}, \ldots, \{label_n\}$ fields. $\{label_1\}$ represents $\{text_1\}, \ldots, \{label_n\}$ represents $\{text_n\}$. 3. The length is approximately $\{m\}$ tokens. 4. Maintain a similar legal document writing style, requiring rigor, no errors, and conformity to actual conditions. 5. Pick a random date in the past.

Table 8: Prompts of data augmentation for tail categories. In practice, $\{document\}$ is replaced with tail samples. $\{label_n\}$ represents a label of $\{document\}$, where n is the number of assigned labels. The corresponding content $\{text_n\}$, expanded through the label semantic enhancement module, is added to the prompt to enrich the information and enhance its expressiveness. $\{m\}$ specifies the desired sample length.