

RESEARCH ARTICLE

MLGN:A Multi-Label Guided Network for Improving Text Classification

QIANG LIU¹, JINGZHE CHEN², FAN CHEN¹, KEJIE FANG¹, PENG AN³,
YIMING ZHANG⁴, AND SHIYU DU^{4,5,6}

¹Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China

²Zhejiang Tianyan Technology Company Ltd., Hangzhou 311215, China

³College of Electronics and Information Engineering, Ningbo University of Technology, Ningbo, Zhejiang 315211, China

⁴Engineering Laboratory of Advanced Energy Materials, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo 315201, China

⁵School of Materials Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China

⁶School of Computer Science, China University of Petroleum (East China), Qingdao 266580, China

Corresponding authors: Shiyu Du (dushiyu@nimte.ac.cn) and Yiming Zhang (ymzhang@nimte.ac.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 52250005 and Grant 21875271, in part by the Key Research and Development Projects of Zhejiang Province under Grant 2022C01236, in part by the Zhejiang Province Key Research and Development Program under Grant 2019C01060, and in part by the Project of the Key Technology for Virtue Reactors from NPIC Entrepreneurship Program of Foshan National Hi-Tech Industrial Development Zone.

ABSTRACT Within natural language processing, multi-label classification is an important but challenging task. It is more complex than single-label classification since the document representations need to cover fine-grained label information, while the labels predicted by the model are often related. Recently, large pre-trained language models have achieved great performance on multi-label classification tasks, typically using embedding of [CLS] vector as the semantic representation of entire document and matching it with candidate labels. However, existing methods tend to ignore label semantics, and the relationships between labels and documents are not effectively mined. In addition, the linear layers used for fine-tuning do not take the correlations between labels into account. In this work, we propose a Multi-Label Guided Network (MLGN) capable to guide document representation with multi-label semantic information. Furthermore, we utilize correlation knowledge to enhance the original label prediction in downstream tasks. The extensive experimental trials show that MLGN transcends previous works on several publicly available datasets. Our source code is available at <https://github.com/L199Q/MLGN>.

INDEX TERMS Multi-label text classification, document representation, label semantics, contrastive learning, label correlation.

I. INTRODUCTION

Multi-label text classification [1], [2] is one of the fundamental tasks in natural language processing (NLP) with a wide range of applications [3], [4]. In text classification, a document can belong to multiple topics and be labeled with multiple tags. For example, in an NLP application in the field of metal materials, the document can be associated with tags such as “computer”, “knowledge graph”, and “materials science”. In fine-grained sentiment analysis, a negative label can be further refined into labels such as

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés .

“depression”, “anger”, “pain”, and “fear”. Multi-label text classification can provide a more refined and comprehensive representation of text content, which is better aligned with the practical needs of the real world and has gradually become the mainstream research direction in text classification.

Multi-label text classification can be classified into two categories: traditional machine learning methods and deep learning methods. To classify text, traditional machine learning methods require text features, which are usually extracted by counting word frequency or using bag-of-words features. These features are then used as input for a classifier, such as decision trees [5], Bayesian [6], SVM [7], or KNN [8], [9]. However, traditional machine learning methods often have

limitations in representing text, as they use discrete representations that are high-dimensional and sparse, and don't capture the semantic relationships between text sequences or contexts. This can make it difficult to learn the meaning of the text, which leads to limited representation ability when building classifier models.

One of the core problems in multi-label text classification is to learn good representations for each input document. In order to capture the semantic features of documents related to labels, research in multi-label text classification has gradually turned to deep learning methods, benefiting from the more powerful ability of deep learning models in text representation and complex feature extraction, as well as further improve the accuracy of text classification. CNN-based deep learning methods have the ability of representation learning [10], [11], but the fixed-size convolutional kernels can only extract local document features and cannot focus on the semantic information of the document context. RNN treats the text as a sequence of words [12], [13], which can capture the correlations between words and learn contextual features. However, due to the sequential processing of text, the computational cost increases with the length of the sentence, which is not instrumental to solving long text problems. Owing to the superiority of Transformer [14] in semantic feature extraction, Google proposed a pre-trained language model called BERT [15], which is based on the encoder structure of Transformer. The model is trained through MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) tasks to obtain excellent text representations, and has achieved the best performance in various natural language processing tasks. This has led to widespread research on pre-trained models by scholars [16], [17]. However, most of them only focus on the feature representations of documents, without explicitly establishing a connection between documents and labels.

In recent years, the semantic information of labels has attracted great attentions of scholars. Guo et al. [18] replaced the original one-hot label encoding with better label distribution generated by label semantic information to improve the final classification performance. Xiong et al. [19] improved the performance of BERT in text classification by utilizing label semantic information. However, they have limitations on the total number of labels, and cannot effectively mine the potential connections between labels and text in large-scale multi-label text classification tasks. To address this issue, HGCLR [20] used the label hierarchy using Graphormer [21], and fused label information with document information to obtain positive samples. GUDN [22] used label semantics to help BERT extract high-quality document features. However, their downstream task fine-tuning only uses a single fully connected layer and does not consider the correlation between labels.

Considering the semantic information of multi-label is not fully utilized to enhance document representation, we propose the LabelInfo module, which employs BERT to extract label semantic information to guide document encoding.

Since the documents and labels of each instance are corresponding, they are considered to be close in the latent space. Therefore, by using contrastive learning to reduce the distance between document and label vectors, we obtain high-quality document representations to improve classification accuracy. In order to effectively utilize label relevance, we propose a new network architecture called LabelNet. The LabelNet architecture is used as an additional enhancement module for existing multi-label text classification architectures to form a new end-to-end model. Our work is summarized as follows:

- 1) The LabelInfo module we proposed combines BERT and contrastive learning loss function of document-label pairs to extract features, which could further guide document representation while obtaining label semantic information, and more effectively to find the latent space between text and labels.
- 2) We propose the LabelNet module, which enhances the original label prediction by utilizing relevance knowledge. This module obtains label relevance predictions by deeply exploring the potential connections between related labels.
- 3) We fuse the LabelInfo and LabelNet modules to propose an end-to-end multi-label guidance network called MLGN. We conducted experiments on two benchmark datasets and achieved state-of-the-art (SOTA) results. Our results demonstrate that MLGN is helpful for multi-label text classification tasks.

II. RELATED WORK

A. DOCUMENT REPRESENTATION

XMLCNN [11] uses dynamic pooling to pool each feature map into multiple features before concatenating them to obtain the document representation. AttentionXML [13] captures long-distance dependencies between words using BiLSTM and uses a multi-label attention mechanism to capture the most relevant parts of the text for each label. LightXML [23] integrates BERT, RoBERTa [16], and XLNet [17] models and concatenates the [cls] vectors from the last 5 layers as the text representation. Zhang et al. [24] argue that a global feature vector may not be sufficient to represent semantic information at different levels of granularity in a document and propose to use word-level local features to supplement it for additional gains.

B. LABEL SEMANTICS

Guo et al. [18] address the problem of one-hot encoding for most text classification labels, which ignores the relationships among text, labels and the semantic information of labels, by integrating document representation information into label representation. LSAN [25] constructs label-specific document representation using label semantic information. Xiong et al. [19] use label embedding technology to improve the performance of BERT in text classification. The model trains the document and label information together and

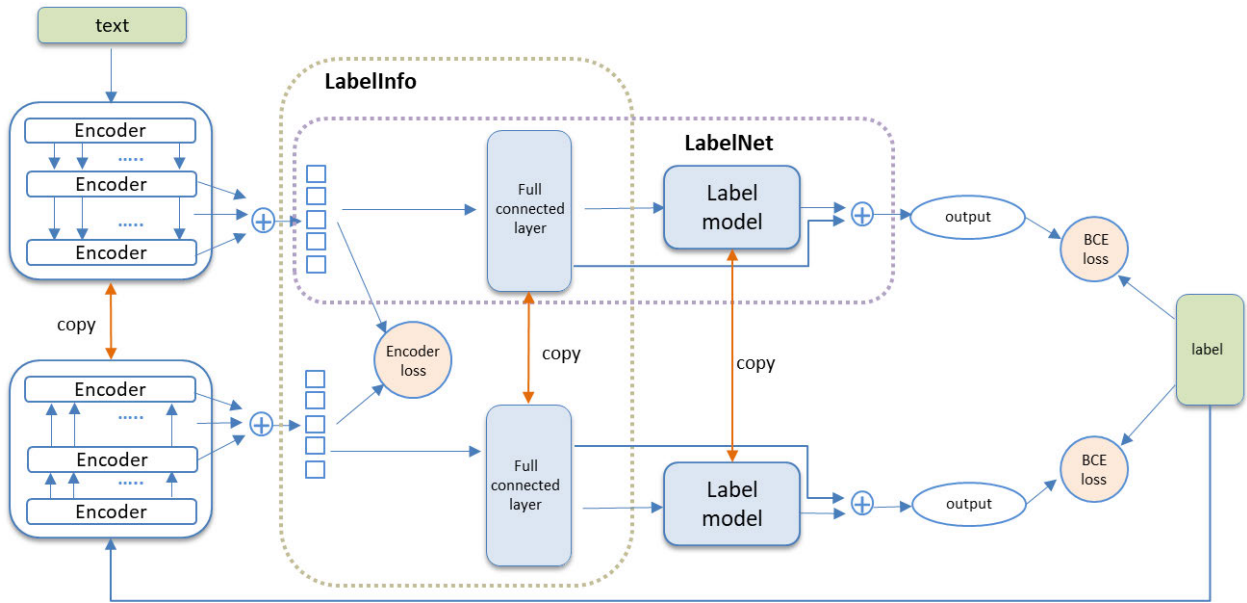


FIGURE 1. The overall framework of the MLGN network.

achieves good results while maintaining almost the same computational cost. GUDN [22] introduces label semantic information to help fine-tune pre-training language models.

C. LABEL CORRELATION

Seq2Seq architecture transforms MLTC into a label sequence generation problem by encoding the input text sequence and decoding the label sequence [26], [27]. This method heavily relies on the predefined label order and is sensitive to the order of the labels [28]. Cornet [29] obtains label relevance by using a linear transformation layer to connect the predicted text results with their original output. However, the linear transformation layer is not enough to deeply explore the internal connections between related labels.

D. CONTRASTIVE LEARNING

In recent years, contrastive learning has been widely used in NLP tasks. Many works have applied it to pre-training language models [30], [31]. SimCSE [32] uses dropout as data augmentation to obtain positive samples and improves the ability of sentence representation. Su et al. [33] help the model obtain text representation that is closer to samples with similar labels by using contrastive learning, thereby improving the quality of KNN retrieval. Suresh et al. [34] propose a label-aware contrastive loss function for fine-grained text classification.

III. PROPOSED METHOD

This work proposes a multi-label guided network (MLGN) that consists of two main modules: the LabelInfo module, which guides document representation with label information, and the LabelNet module, which uses correlation knowledge to enhance original label predictions. Firstly, we use

BERT as a feature extractor to obtain the semantic features of both the document and labels. These semantic features are then input into the LabelInfo module, where label information is treated as positive samples of document information. By using contrastive learning, we can fully explore the relationships between the document and labels. Secondly, the original label prediction output from LabelInfo is used as the input to the Labelmodel, where the combination of the two parts is called LabelNet. This module mainly obtains relevant combinations of original tag predictions by training multiple weight matrices. Finally, MLGN utilizes label semantic information and label correlation to achieve accurate multi-label classification results. The overall framework of this network is shown in Figure 1.

A. PRELIMINARIES

Let Dataset = $\{(x_i, y_i)\}_{i=1}^N$, where x_i is the original document and the i -th text is represented as $x_i = \{w_1, w_2, \dots, w_T\}$, T represents the input length of the document, w_i is the i -th word of the document. $y_i \in \{0, 1\}^L$ is the corresponding label set for x_i , L is the label set of the dataset, and N is the total number of examples in the dataset. The classifier computes the probability p_i of each label being true, where $p_i = \{p_1, p_2, \dots, p_L\}$. The binary cross entropy (BCE) loss between p_i and y_i is calculated as follows:

$$L_{BCE}(p_i, y_i) = -\frac{1}{L} \sum_{l \in L} [y_l \log p_l + (1 - y_l) \log (1 - p_l)]. \quad (1)$$

Before being input to the BERT model, the document x_i is typically prepended with a special [CLS] token. For a Transformer model with n layers, the hidden representations

of the n-th layer is denoted as:

$$\phi_{\text{bert}}^{(n)}(x_i) = \{h_{\text{cls}}^{(n)}, h_1^{(n)}, \dots, h_T^{(n)}\}. \quad (2)$$

B. LabelInfo

The [CLS] token in BERT utilizes a self-attention mechanism to obtain sentence-level information representation, which can capture contextual information representations in different contexts. We use the embedded [CLS] in BERT as the semantic feature of the document and label. However, using only the [CLS] of the last layer is insufficient for exploring the relationships between the semantic features of the document and the label. This is mainly because there are issues with obtaining the feature information of the label: (1) The label information is composed of related words and does not contain contextual semantic information. (2) The label information is usually shorter than document information and is only provided by multiple labels corresponding to each document. To address these issues, we concatenate the [CLS] of the last five layers of BERT as its feature representation to obtain hierarchically rich semantic information. Additionally, the label and text share the same BERT, which significantly reduces the model size and complexity, accelerating convergence. During the training phase, document features and label features are asynchronously extracted. The text feature expression is as follows:

$$h = \text{concat} \left(h_{\text{cls}}^{(-1)}, h_{\text{cls}}^{(-2)}, h_{\text{cls}}^{(-3)}, h_{\text{cls}}^{(-4)}, h_{\text{cls}}^{(-5)} \right). \quad (3)$$

Simply relying on a simple fully connected layer to link the semantic information of the text and the label to the one-hot label is unstable and uncertain. An effective and simple method to solve this problem is to create an association mechanism between the label information and the document information. Therefore, we propose a document-label contrastive learning loss function to solve the above problem. We guide the encoding of K document features, thus using the features of K multi-label sets associated with the document, so the total number of samples is 2K, $I = \{1, \dots, 2K\}$. We represent the index of the multi-label set corresponding to the i-th document as the label(i), and the negative sample is the remaining sample in I. The document-label contrastive learning loss function expression is as follows:

$$L_{\text{Encoder}} = \sum_{i=1}^{2k} -\log \frac{\exp(h_i \cdot h_{\text{label}(i)}/\tau)}{\sum_{k \in I/i} \exp(h_i \cdot h_k/\tau)}. \quad (4)$$

where τ is the temperature coefficient, which helps to better distinguish between positive and negative samples. h_i is the concatenated vector of the last 5 layers of [CLS] obtained from h to obtain x_i , which is the normalized document representation vector.

The success of LabelInfo can be attributed to two main factors. First, it utilizes the semantic information of the label to guide BERT in extracting features related to the label from the document information, resulting in a document representation with label information guidance. Second, it creates

an association mechanism between the semantic information of the label and the document information, using the label feature encoding as a positive sample of the document feature. Through the document-label contrastive learning loss function, it brings the projection space distance between the document feature and the associated label feature closer, while also guiding dissimilar instances to distance themselves further apart in the projection space.

C. LabelNet

LabelNet is composed of raw label predictions and Labelmodel, which is a computational unit that maps the raw label predictions to enhanced label predictions based on label correlations. The building blocks of LabelNet are shown in Figure 2. Formally, the LabelNet construction is defined as:

$$Y = y + \text{Labelmodel}(y). \quad (5)$$

where Y and y are the output and input of the LabelNet module. Specifically, y is the raw label prediction before LabelNet, and Y is the enhanced label prediction with correlations learned by the Label model.

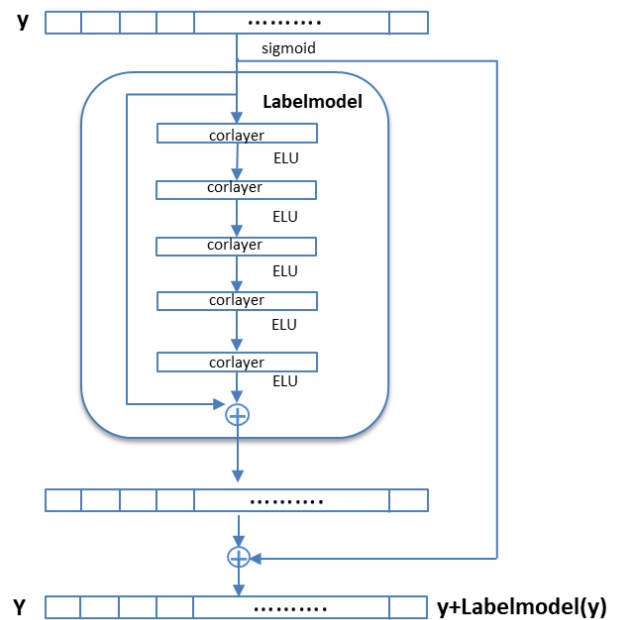


FIGURE 2. The framework of the LabelNet network.

The simplest design for the label correlation module is to add a linear layer after y, similar to Cornet [19], to obtain the correlated label prediction through additional weight training. However, a single linear layer is insufficient to deepen the connection between labels, and the learned label correlations are shallow. Therefore, we propose to deepen the label correlation layer, where the output of the previous layer serves as the input of the next layer. This allows us to focus on the correlation between related labels and improve the learned label correlation. However, deepening the network layers may cause the predicted results to diverge. To address this issue, we use the original label prediction as a constraint

TABLE 1. Summary of experimental datasets.

| Datasets | N_{train} | N_{test} | D | L | \bar{L} | \tilde{L} | \bar{w}_{train} | \bar{w}_{test} |
|----------|--------------------|-------------------|--------|------|-----------|-------------|--------------------------|-------------------------|
| EURLex | 15499 | 3865 | 186104 | 3956 | 5.30 | 20.79 | 1248.58 | 1230.40 |
| AAPD | 54840 | 1000 | 69399 | 54 | 2.41 | 2444.04 | 163.42 | 171.65 |

Notes: N_{train} is the number of training instances, N_{test} is the number of test instances, D is the total number of words, L is the total number of classes, \bar{L} is the average number of documents per document, \tilde{L} is the average number of documents per label, \bar{w}_{train} is the average number of words per document in the training set, \bar{w}_{test} is the average number of words per document in the testing set.

for label correlation learning through residual connections. In multi-label text classification, the total number of labels can be large, but only a few of them are typically related. The majority of labels are often unrelated to each other. Therefore, the length of our label correlation layer can be much smaller than the total length of the labels, which not only allows us to focus on learning deep correlations but also reduces the model parameters and accelerates model convergence. The label correlation layer is defined as:

$$\text{corlayer}(v) = \delta(Wv + b). \quad (6)$$

where v is the input of the label correlation layer, W and b are the weight matrix and bias of the label correlation layer, and δ is the ELU activation function.

Considering that the raw label prediction, as the output of labelInfo, has already achieved high accuracy and plays an important guiding role in the label correlation prediction as the input of LabelNet. We again amplify the effect of the original label prediction through residual connections.

D. TRAINING AND PREDICTION

1) TRAINING

The MLGN is an end-to-end multi-label classification model that is made possible by the constructed LabelInfo and LabelNet. The goal of MLGN is to minimize the target loss function L_{sum} , which includes L_{Encoder} , $L_{\text{BCE}}(p_i^{\text{text}}, y_i)$, and $L_{\text{BCE}}(p_i^{\text{label}}, y_i)$. The specific formula is defined as:

$$L_{\text{sum}} = \lambda L_{\text{Encoder}} + L_{\text{BCE}}(p_i^{\text{text}}, y_i) + L_{\text{BCE}}(p_i^{\text{label}}, y_i). \quad (7)$$

Here, λ is an adjustable coefficient for the contrastive learning loss function of document-label, which is used to control the balance between the losses. p_i^{text} is the final probability of the i -th document's semantic information as input, and p_i^{label} is the final probability of the semantic information of the label set associated with the i -th document as input.

2) PREDICTION

As we fully utilize the semantic information of the labels during the training phase, and MLGN is trained when L_{sum} is minimized. Even in the prediction phase without the condition of label semantic information, the MLGN model that relies only on the semantic information of the document can

obtain guidance for the label semantic information during document representation and final prediction, and the final prediction probability of the i -th document is p_i^{text} .

IV. EXPERIMENTS

A. DATASETS

- 1) EURLex dataset [35] is a collection of documents related to EU law, containing 3956 topics with 15449 documents in the training set and 3865 in the test set.
- 2) AAPD dataset [27] consists of 55840 abstracts and corresponding topics from papers in the computer science field on arXiv, with 54840 in the training set and 1000 in the test set.

For both datasets, the maximum document length is 512, and if the number of words in a document exceeds 512, we truncate the document to the maximum number of words. All methods are trained and tested on the datasets summarized in Table 1.

B. EVALUATION METRICS

Considering that in multi-label text classification datasets, the number of labels per sample is sparse with respect to the total set of labels. For this reason in the evaluation phase, we provide a short ranked list of potentially relevant labels for each instance and evaluate the quality of these ranked lists, focusing on the scores at the top of each list. Therefore, this study uses two evaluation metrics to verify the validity of MLGN: the top k precision ($p@k$) and the normalized discounted cumulative gain ($nDCG@k$). We calculate $p@k$ and $nDCG@k$ using the following equations:

$$p@k = \frac{1}{k} \sum_{t \in \text{rank}_k(\hat{y})} y_t. \quad (8)$$

$$DCG@k = \frac{1}{k} \sum_{t \in \text{rank}_k(\hat{y})} \frac{y_t}{\log(t+1)}. \quad (9)$$

$$IDCG@k = \frac{1}{\sum_{t=1}^{\min(k, \|y\|_0)} \frac{1}{\log(t+1)}}. \quad (10)$$

$$nDCG@k = \frac{DCG@k}{IDCG@k}. \quad (11)$$

where \hat{y} is the predicted score vector, $\text{rank}_k(\hat{y})$ is the indices of the top k scores in \hat{y} , y is the true label vector, $y \in \{0, 1\}^L$,

and $\|y\|_0$ is the number of relevant labels in y , that is, the number of 1 in y .

C. BASELINE MODELS

To fully validate the effectiveness of the MLGN model, this work compares it with state-of-the-art models for multi-label text classification tasks in recent years, with parameters either adopted from their original papers or determined through experiments. The baseline models are as follows:

- 1) XMLCNN [11]: This model uses a convolutional neural network to represent text and dynamically extracts hierarchically rich semantic features from text using dynamic pooling.
- 2) AttentionXML [13]: A model based on label trees that leverages the advantages of BiLSTM networks to obtain contextual semantic information and obtains document representation related to label information through a label attention mechanism.
- 3) CornetAttentionXML [29]: An architecture that uses AttentionXML as the text encoder and can leverage the correlation information between different labels by attaching a Cornet network module.
- 4) LightXML [23]: A lightweight deep framework with dynamic negative label sampling. To ensure experimental fairness, we use a BERT model to reproduce it.
- 5) GUDN [22]: A multi-label classification model that utilizes label semantic information, which is the most similar work to MLGN because they both use label semantic information as guidance. However, it does not effectively establish a mechanism for the association between labels and documents, and lacks work on label correlation.

D. PARAMETER SETTINGS

All of our experiments were conducted on a computer with a Tesla V100 GPU and Intel Xeon 4210 (2.4G, 10C) CPU. Our models used the pre-trained bert-base-uncased version as a feature extractor, which consists of 12 Transformer blocks with 12 self-attention heads and a hidden size of 768. The text representation had a dropout rate of 0.5, and the adjustable coefficient was set to 0.01. For the EURlex dataset, we set the learning rate to $5e-5$, the label Encoder input to the semantic information of multiple labels corresponding to each sample, the temperature coefficient was set to 5, the dimension of the label correlation layer in LabelNet was set to 600, and the number of training rounds was set to 40, with the training batch size of 8 and the testing batch size of 16. For the AAPD dataset, we set the learning rate to $5e-6$, the label Encoder input to the augmented label semantics (arXiv labels with complete semantics as shown in Table 2) + the first 48 words of the document, the temperature coefficient was set to 1, the dimension of the label correlation layer in LabelNet was set to 30, and the number of training rounds was set to 20, with the training batch size of 16 and the testing batch size of 16.

TABLE 2. Labels and enhanced labels.

| Labels | Enhanced Labels |
|--------------------|---|
| cmp-lg | cmp-lg |
| cond-mat.dis-nn | Condensed Matter Disordered Systems Neural Networks |
| cond-mat.stat-mech | Condensed Matter Statistical Mechanics |
| cs.AI | Computer Artificial Intelligence |
| cs.CC | Computer Computational Complexity |
| cs.CE | Computer Computational Engineering Finance Science |
| cs.CG | Computer Computational Geometry |
| cs.CL | Computer Computation Language |
| cs.CR | Computer Cryptography Security |
| cs.CV | Computer Vision Pattern Recognition |
| cs.CY | Computers Society |
| cs.DB | Computer Databases |
| cs.DC | Computer Distributed Parallel Cluster Computing |
| cs.DL | Computer Digital Libraries |
| cs.DM | Computer Discrete Mathematics |
| cs.DS | Computer Data Structures Algorithms |
| cs.FL | Computer Formal Languages Automata Theory |
| cs.GT | Computer Science Game Theory |
| cs.HC | Human-Computer Interaction |
| cs.IR | Computer Information Retrieval |
| cs.IT | Computer Information Theory |
| cs.LG | Computer Machine Learning |
| cs.LO | Logic in Computer Science |
| cs.MA | Computer Multiagent Systems |
| cs.MM | Computer Multimedia |
| cs.MS | Computer Mathematical Software |
| cs.NA | Computer Numerical Analysis |
| cs.NE | Computer eural Evolutionary Computing |
| cs.NI | Computer Networking Internet Architecture |
| cs.PF | Computer Performance |
| cs.PL | Computer Programming Languages |
| cs.RO | Computer Robotics |
| cs.SC | Computer Symbolic Computation |
| cs.SE | Computer Software Engineering |
| cs.SI | Computer Social Information Networks |
| cs.SY | Computer Systems Control |
| math.CO | Mathematics Combinatorics |
| math.IT | Mathematics Information Theory |
| math.LO | Mathematics Logic |
| math.NA | Mathematics Numerical Analysis |
| math.NT | Mathematics Number Theory |
| math.OC | Mathematics Optimization Control |
| math.PR | Mathematics Probability |
| math.ST | Mathematics Statistics Theory |
| nlin.AO | NMathematics nonlinear Adaptation Self-Organizing Systems |
| physics.data-an | Physics Data Analysis Statistics Probability |
| physics.soc-ph | Physics Society |
| q-bio.NC | Quantitative Biology Neurons Cognition |
| q-bio.QM | Biology Quantitative Methods |
| quant-ph | Quantum Physics |
| stat.AP | Statistics Applications |
| stat.ME | Statistics Methodology |
| stat.ML | Statistics Machine Learning |
| stat.TH | Statistics Theory |

Notes: The label "cmp-lg" was not found on arXiv, so the enhanced label remains the same as the original label.

E. PERFORMANCE COMPARISON

We evaluated the performance advantage of our proposed MLGN model over existing models using the $p@k$ and $nDCG@k$ evaluation metrics. The best results are shown in bold in Table 3. It can be seen that our proposed MLGN model transcends current state-of-the-art multi-label text classification models on every metric. However, XMLCNN performs the worst, mainly because the text representation obtained through CNN lacks contextual semantic relationships. AttentionXML improves upon XMLCNN with the attention mechanism, but it only focuses on document representation and does not consider the work of label relevance. This is why CornetAttentionXML surpasses it. The Transformer model relies on its powerful feature extraction ability to obtain high-quality text representations, and its

TABLE 3. Comparison of different models.

| Datasets | | XMLCNN | AttentionXML | CornetAttentionXML | LightXML | GUDN | MLGN |
|----------|----------|--------|--------------|--------------------|----------|-------|--------------|
| EURLex | $P@1$ | 76.81 | 85.90 | 85.85 | 86.03 | 85.51 | 86.31 |
| | $P@3$ | 62.79 | 73.01 | 73.32 | 74.19 | 74.10 | 74.77 |
| | $P@5$ | 51.56 | 61.00 | 61.68 | 62.27 | 62.14 | 62.66 |
| | $nDCG@3$ | 66.44 | 76.41 | 76.61 | 77.41 | 77.23 | 77.97 |
| | $nDCG@5$ | 60.47 | 70.47 | 70.94 | 71.70 | 71.49 | 72.13 |
| | | | | | | | |
| AAPD | $P@1$ | 74.38 | 83.70 | 85.00 | 85.80 | 85.80 | 86.10 |
| | $P@3$ | 53.84 | 60.63 | 61.57 | 61.30 | 62.30 | 62.57 |
| | $P@5$ | 37.79 | 41.64 | 41.76 | 42.02 | 42.42 | 42.44 |
| | $nDCG@3$ | 71.12 | 79.90 | 81.25 | 81.26 | 81.97 | 82.45 |
| | $nDCG@5$ | 75.93 | 84.10 | 84.90 | 85.32 | 85.87 | 86.17 |
| | | | | | | | |

Notes: We compared our model with state-of-the-art multi-label text classification models. Note that to ensure experimental fairness, we do not use model ensemble. For the LightXML model, we selected the best experimental results from BERT, RoBERTa, and XLNet to display. For AAPD, to ensure the effectiveness of label correlation, we chose a dimension of 30 (<54) for the Cornet module.

classification performance is generally better than that of traditional neural networks. GUDN enhances document representation with label semantic information and performs better than LightXML on the AAPD dataset. However, it is inferior on the EURLex dataset as the documents in this dataset are not traditional English words and do not fully explore the relationships between labels and documents.

The MLGN successfully overcomes the limitations of the previously mentioned models for three main reasons: (1) it employs pre-trained language models for complex feature extraction, obtaining rich text representations through the [CLS] vector of the last 5 layers; (2) it utilizes label semantic information to enhance document representation and fully capitalizes on the contrastive learning loss function of document-label pairs to establish a connection between the potential space of documents and labels; (3) it takes label relevance into account for downstream tasks by utilizing the depth of the label relevance layer to improve original label prediction and capture relevant knowledge.

F. ABLATION EXPERIMENTS

In this section, we evaluated the efficacy of LabelInfo and LabelNet in the MLGN by comparing the performance of a single BERT module, a BERT+LabelInfo module, and a BERT+LabelNet module. To ensure the fairness of the experiments, we used the [CLS] vector of the last 5 layers of BERT as the text representation for all three modules. Notably, the input to the single BERT module was the document, the input to the BERT+LabelInfo module was the document plus the semantic information of the label, and the input to the BERT+LabelNet module was the document only. The impact of different modules is shown in Tables 4 and 5.

In the LabelInfo module, the BERT+LabelInfo module outperformed the single BERT module on both the $p@k$ and $nDCG@k$ metrics on both datasets, which fully demonstrated the feasibility of using label semantic information to guide document representation. Moreover, to investigate the impact of label information input in LabelInfo on model performance, we found that the LabelInfo module had a greater

TABLE 4. Comparison of the ablation results of each module on the EURLex dataset.

| Modules | $p@1$ | $p@3$ | $p@5$ | $nDCG@3$ | $nDCG@5$ |
|----------------|-------|-------|-------|----------|----------|
| BERT | 84.71 | 73.44 | 61.82 | 76.55 | 71.03 |
| BERT+LabelNet | 85.41 | 74.03 | 61.95 | 77.10 | 71.28 |
| BERT+LabelInfo | 85.41 | 74.29 | 62.37 | 77.34 | 71.66 |

TABLE 5. Comparison of the ablation results of each module on the AAPD dataset.

| Modules | $p@1$ | $p@3$ | $p@5$ | $nDCG@3$ | $nDCG@5$ |
|----------------|-------|-------|-------|----------|----------|
| BERT | 85.40 | 61.63 | 42.02 | 81.50 | 85.45 |
| BERT+LabelNet | 85.70 | 61.77 | 42.12 | 81.60 | 85.53 |
| BERT+LabelInfo | 85.40 | 61.97 | 42.28 | 81.73 | 85.63 |

performance improvement on the EURLex dataset than on the AAPD dataset. This can be attributed to the fact that the EURLex dataset has a higher average number of labels per sample than the AAPD dataset (as shown in Table 1), and the semantic information of the labels in the EURLex dataset is more abundant, while the semantic information of the labels in the AAPD dataset needs to be obtained from external knowledge (as shown in Table 2). This finding highlights the direct impact of the quantity and quality of label information on improving model performance.

In the LabelNet module, the performance of the single BERT module was inferior to that of the BERT+LabelNet module on both the EURLex dataset and the AAPD dataset, which demonstrated the feasibility of exploring the relevance of original label prediction. This can be attributed to the fact that the single BERT module, which fine-tunes the last 5 layers' [CLS] vectors using a linear layer to obtain label probabilities in downstream tasks, often ignores the relevance between labels. However, the LabelNet module can solve this problem by learning the potential connections between labels. This is particularly evident on the EURLex dataset, which has a large number of total labels (3956 in total), most of which are often unrelated. LabelNet can exclude irrelevant label interference by obtaining a few relevant labels to improve the model's performance.

G. DOCUMENT REPRESENTATION EXPERIMENTS

This section presents further experiments to explore the effectiveness of the contrastive learning loss function for document representations in comparison to label information, as shown in Table 6. The MLGN network with a contrastive learning loss function that incorporates document-label pairs demonstrated the best performance on the nDCG@3 and nDCG@5 metrics, confirming the effectiveness of the contrastive learning loss function in fully exploring the potential space between document and label semantics. This successfully establishes a bridge between the semantic information of documents and labels, enabling label information to effectively guide and enhance document representations.

TABLE 6. Comparison of the role of the associated mechanism based on the document-label contrastive learning loss function.

| Datasets | | MLGN | MLGN/ $L_{Encoder}$ |
|----------|----------|--------------|---------------------|
| EURLex | $nDCG@3$ | 77.97 | 77.68 |
| | $nDCG@5$ | 72.13 | 71.93 |
| AAPD | $nDCG@3$ | 82.45 | 82.33 |
| | $nDCG@5$ | 86.17 | 86.11 |

Notes: $MLGN/L_{Encoder}$ represents the MLGN model without an associated mechanism.

The MLGN utilizes the [CLS] vectors of the last 5 layers of BERT and concatenates them to obtain feature representations, aiming to obtain higher-level and more hierarchically rich semantic features to improve the effectiveness of the model. To analyze how multi-layer text representations affect model performance, we compared it with only using the last layer’s [CLS] vector as the text representation on the EURLex dataset. The experimental results are shown in Figure 3, where the multi-layer [CLS] has higher accuracy, indicating that multi-layer text representations can extract higher-level semantic features, and multiple layers can achieve the same performance as a single layer using less time. For the final metrics, multi-layers improved p@5 and nDCG@5 by 0.2% and 0.35%, respectively.

To further explore the interpretability of MLGN’s high-quality document representations, some visual work was done on the document representations of MLGN in this section. As shown in Figure 4, the document representations in BERT are relatively scattered within each class and close between classes, while in the MLGN model, each point within each class is more concentrated, especially at the boundaries of cs.DS and cs.LG. Although the points of cs.LG appear at the boundary, they still appear in a concentrated manner. Additionally, cs.DS and cs.LO are separated by a relatively long distance in MLGN, while BERT does not distinguish between these two classes well, with a close and relatively overlapping distance.

H. LABEL CORRELATION EXPERIMENTS

In this section, we compare the effectiveness of our label correlation module, BERT+LabelNet, with BERT+Cornet.

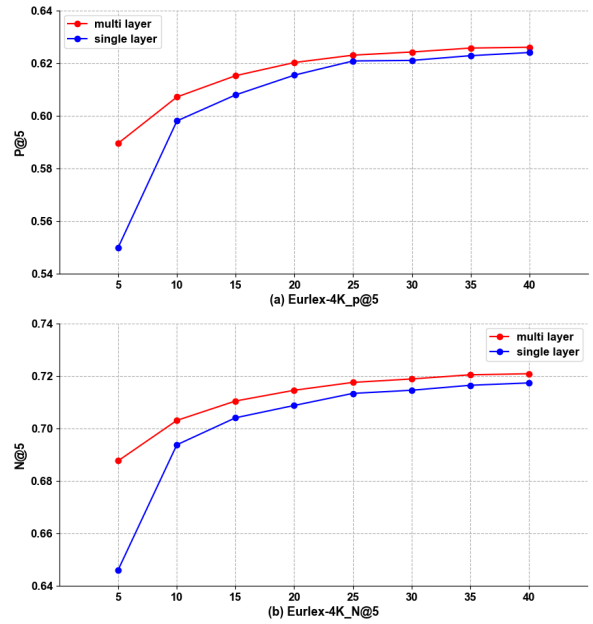


FIGURE 3. Effect of multi-layer text representations. Multi-layers concatenate the [CLS] vectors of the last 5 layers of BERT as the text representation, and single layer only uses the last layer’s [CLS] vector as the text representation.

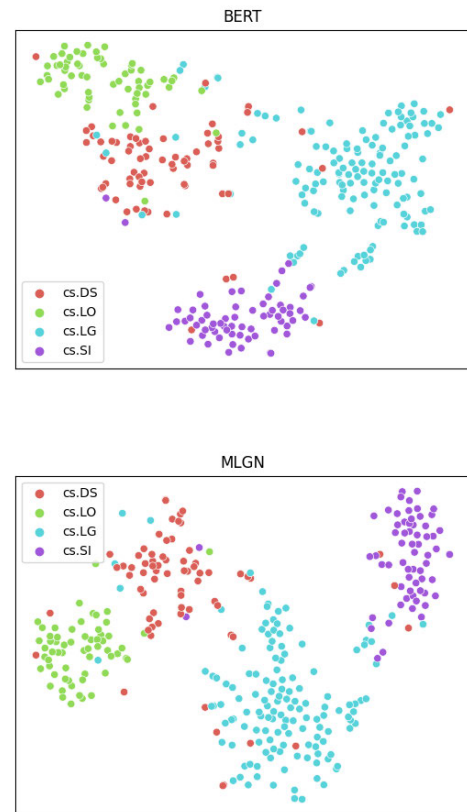


FIGURE 4. tSNE visualization of document representation vectors on the AAPD test set. The document representation is obtained by concatenating the last 5 layers’ [CLS] vectors in bert, and the tSNE maps learned on the cs.DS, cs.LO, cs.LG, and cs.SI categories are shown.

The experimental results are shown in Figure 5. We propose the BERT+LabelNet module, which outperforms the

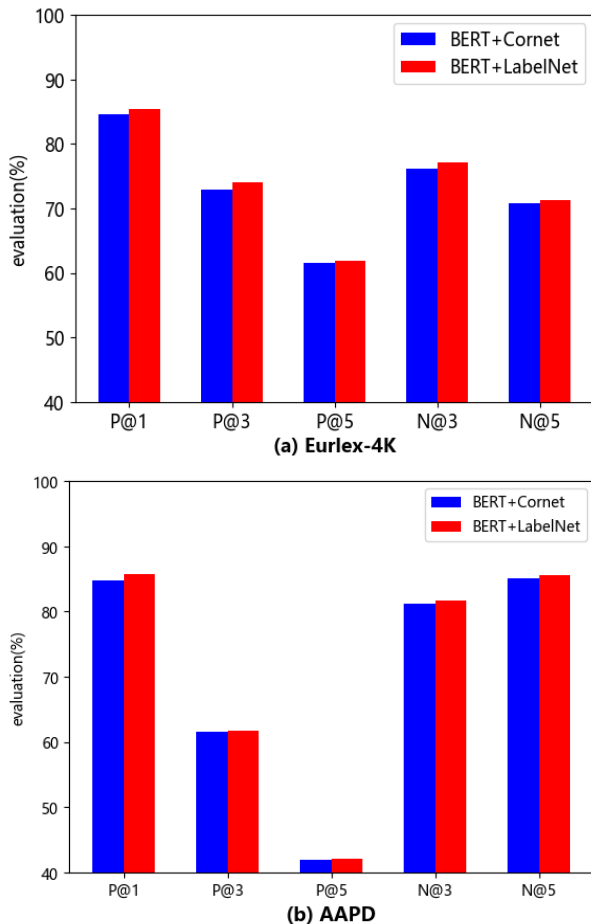


FIGURE 5. Comparison of the LabelNet and Cornet modules.

TABLE 7. Prediction scores of related labels using MLGN on the EURLex test set.

| Label | BERT | MLGN |
|------------------------|--------|--------|
| "sugar" | 0.9759 | 0.9996 |
| "sugar_product" | 0.5353 | 0.9610 |
| "community_statistics" | 0.9985 | 0.9999 |
| "statistical_method" | 0.2280 | 0.8093 |
| "animal_disease" | 0.9999 | 0.9999 |
| "poultry" | 0.9925 | 0.9934 |
| "slaughter_of_animals" | 0.9127 | 0.9912 |

Notes: Three samples with related labels are selected, and their predicted scores are shown.

BERT+Cornet module on both datasets. This indicates that deepening the label correlation layers and exploring the potential connections between related labels can help improve the prediction of original labels and enhance their correlation.

To further investigate how MLGN leverages correlation knowledge to enhance the prediction of original labels, we present the prediction scores of several related labels on the EURLex test set. As shown in Table 7, MLGN has higher

scores than a single BERT model, and the scores are closer in each instance. This indicates that MLGN can identify related labels and improve model performance.

V. CONCLUSION

This study proposes a multi-label guided network called MLGN that incorporates label semantic information to enhance document representation and improves label relevance through original label prediction. As shown in Table 3, our model outperforms prior work on both the AAPD dataset and the Eurlex dataset on five evaluation metrics including $p@k$ and $nDCG@k$, demonstrating the superiority of MLGN in multi-label text classification tasks. Our ablation and analysis experiments reveal two points: (1) labelinfo can utilize the semantic information of multiple labels to guide document encoding and obtain high-quality document representation. Our proposed document-label contrastive learning loss function can also fully explore the potential space of documents and labels. This solves the problem of existing methods that tend to ignore label semantic information, and fail to effectively extract the relationship between labels and documents. (2) labelnet is feasible in acquiring deep label relevance by deepening the label-related layers. Using the knowledge of relevance to enhance original label prediction addresses the issue that the linear layer used in pre-trained language models for fine-tuning does not consider the relationship between labels.

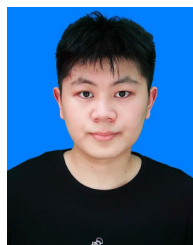
We also demonstrate the feasibility of our approach through label information enhancement on the AAPD dataset, which has a small amount of label semantic information that can be utilized. However, this enhanced label information is achieved by using subject-specific knowledge, which has a certain level of specificity. Additionally, as shown in Table 6, the effect of our document-label contrastive learning loss function on document representation is not significant enough, mainly because the loss function treats each instance's label set as a whole, and cannot identify the internal connections between labels within the label set.

In future work, we aim to explore label information enhancement strategies for predicting labels efficiently and accurately in situations where label semantic information is lacking. We also plan to investigate fine-grained document-label contrastive learning and focus on fully exploring the potential relationships between similar documents through their shared label attributes.

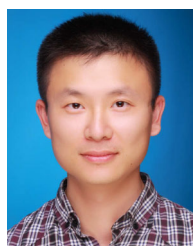
REFERENCES

- [1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, Apr. 2022.
- [2] M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A survey of multi-label classification based on supervised and semi-supervised learning," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 3, pp. 697–724, Mar. 2023.
- [3] T. Kuyucuk and L. Çalli, "Using multi-label classification methods to analyze complaints against cargo services during the COVID-19 outbreak: Comparing survey-based and word-based labeling," *Sakarya Univ. J. Comput. Inf. Sci.*, vol. 5, no. 3, pp. 371–384, Dec. 2022.

- [4] Y. Kementchedjheva and I. Chalkidis, "An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text," 2023, *arXiv:2305.05627*.
- [5] A. Law and A. Ghosh, "Multi-label classification using binary tree of classifiers," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 677–689, Jun. 2022.
- [6] R. Wang, S. Ye, K. Li, and S. Kwong, "Bayesian network based label correlation analysis for multi-label classifier chain," *Inf. Sci.*, vol. 554, pp. 256–275, Apr. 2021.
- [7] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5948–5960, Oct. 2018.
- [8] M. Roseberry, B. Krawczyk, Y. Djenouri, and A. Cano, "Self-adjusting k nearest neighbors for continual learning from multi-label drifting data streams," *Neurocomputing*, vol. 442, pp. 10–25, Jun. 2021.
- [9] G. Alberghini, S. Barbon Jr., and A. Cano, "Adaptive ensemble of self-adjusting nearest neighbor subspaces for multi-label drifting data streams," *Neurocomputing*, vol. 481, pp. 228–248, Apr. 2022.
- [10] Y. Chen, "Convolutional neural network for sentence classification," M.S. thesis, Univ. Waterloo, Waterloo, ON, Canada, 2015.
- [11] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
- [12] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015, pp. 1–7.
- [13] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
- [17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [18] B. Guo, S. Han, X. Han, H. Huang, and T. Lu, "Label confusion learning to enhance text classification models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 14, 2021, pp. 12929–12936.
- [19] Y. Xiong, Y. Feng, H. Wu, H. Kamigaito, and M. Okumura, "Fusing label embedding into BERT: An efficient improvement for text classification," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 1743–1750.
- [20] Z. Wang, P. Wang, L. Huang, X. Sun, and H. Wang, "Incorporating hierarchy into text encoder: A contrastive learning approach for hierarchical text classification," 2022, *arXiv:2203.03825*.
- [21] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform badly for graph representation?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28877–28888.
- [22] Q. Wang, J. Zhu, H. Shu, K. O. Asamoah, J. Shi, and C. Zhou, "GUDN: A novel guide network with label reinforcement strategy for extreme multi-label text classification," 2022, *arXiv:2201.11582*.
- [23] T. Jiang, D. Wang, L. Sun, H. Yang, Z. Zhao, and F. Zhuang, "LightXML: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 9, 2021, pp. 7987–7994.
- [24] R. Zhang, Y.-S. Wang, Y. Yang, T. Vu, and L. Lei, "Exploiting local and global features in transformer-based extreme multi-label text classification," 2022, *arXiv:2204.00933*.
- [25] L. Xiao, X. Huang, B. Chen, and L. Jing, "Label-specific document representation for multi-label text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 466–475.
- [26] J. Nam, E. L. Mencía, H. J. Kim, and J. Fürnkranz, "Maximizing subset accuracy with recurrent neural networks in multi-label classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [27] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: Sequence generation model for multi-label classification," 2018, *arXiv:1806.04822*.
- [28] K. Qin, C. Li, V. Pavlu, and J. A. Aslam, "Adapting RNN sequence prediction model to multi-label set prediction," 2019, *arXiv:1904.05829*.
- [29] G. Xun, K. Jha, J. Sun, and A. Zhang, "Correlation networks for extreme multi-label text classification," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1074–1082.
- [30] T. Kim, K. M. Yoo, and S.-G. Lee, "Self-guided contrastive learning for BERT sentence representations," 2021, *arXiv:2106.07345*.
- [31] L. Pan, C.-W. Hang, A. Sil, and S. Potdar, "Improved text classification via contrastive adversarial training," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, pp. 11130–11138.
- [32] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.
- [33] X. Su, R. Wang, and X. Dai, "Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2022, pp. 672–679.
- [34] V. Suresh and D. C. Ong, "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification," 2021, *arXiv:2109.05427*.
- [35] E. L. Mencía and J. Fürnkranz, "Efficient pairwise multilabel classification for large-scale problems in the legal domain," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Antwerp, Belgium: Springer, Sep. 2008, pp. 50–65.



QIANG LIU received the bachelor's degree in software engineering from the Anhui Institute of Information Technology, in 2020. He is currently pursuing the master's degree in computer technology with Ningbo University. His current research interests include text classification and information extraction.



JINGZHE CHEN received the Ph.D. degree in condensed matter physics from Peking University, in 2008. From 2008 to 2010, he was a Postdoctoral Researcher with the Physics Department, Technical University of Denmark. From 2010 to 2013, he was a Postdoctoral Researcher with the Physics Department, McGill University. He is currently an Associate Professor with the Physics Department, Shanghai University, China. He has published more than 20 articles in *Journal of the American*

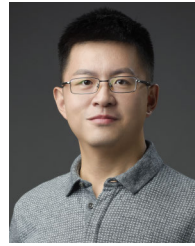
Chemical Society, *Advanced Materials*, *Nanoscale*, *Physical Review B*, and other journals with more than 500 citations. His research interests include the theory of condensed matter physics and computational algorithms in physics and other fields.



FAN CHEN received the bachelor's degree in energy power and engineering from the Changsha University of Technology, in 2021. He is currently pursuing the master's degree in computer technology with Ningbo University. His current research interests include semantic segmentation and semi-supervised learning.



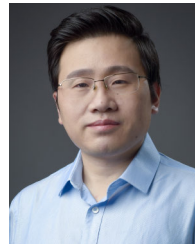
KEJIE FANG received the bachelor's degree in materials science and engineering from Taizhou University, in 2021. He is currently pursuing the master's degree in computer technology with Ningbo University. His current research interests include machine learning and neural networks.



YIMING ZHANG received the master's degree (Hons.) in engineering and the Ph.D. degree from the Queen Mary University of London, in 2006 and 2010, respectively. From 2001 to 2003, he studied mechanical and electronic engineering with the Zhejiang University of Technology. He is currently an Associate Researcher. He is also mainly engaging in the application of material informatics, intelligent optimization algorithms, and system design methods in the research and development of new ternary layered materials and extreme environmental energy materials. He has published more than 40 articles in *Acta Materialia*, *Materialia*, *Philosophical Magazine*, *Journal of Power Sources*, and other magazines, with more than 400 citations, six patents, and five software copyrights. He is a member of the British Society of Minerals, Mining and Materials (IOM3).



PENG AN received the bachelor's and Ph.D. degrees from the Department of Engineering Physics, Tsinghua University, China. From August 2015 to March 2016, he was a Visiting Scientist with Heidelberg University, Germany. He is currently a Professor with the Ningbo University of Technology and a Supervisor of postgraduate students. He has published more than 20 articles in *Information Processing and Management*, *Complexity*, and other magazines, with more than 180 citations, and holds eight patents and two software copyrights. His current research interests include pattern recognition, machine learning, and wireless sensor networks.



SHIYU DU received the Ph.D. degree from the Department of Chemistry, Purdue University, USA, in 2009. From July 2009 to December 2013, he was a Postdoctoral Researcher and a Visiting Scientist with the Los Alamos National Laboratory, USA. During his stay with the Los Alamos National Laboratory, he mainly undertook the theoretical calculation of key physical properties of nuclear fuel and nuclear material. He returned to China working full-time, in January 2014, and established a research team of "Theoretical Design and Performance Simulation of Energy Materials" with the Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, where he is currently a Ph.D. Supervisor and a Professor. He is the Chief Scientist of the National Key Research and Development Program. He is also working in the structural design and characterization of various energy and structural materials, such as nuclear materials by the strategy of materials genome initiative and the multiscale computer simulation method. He has published more than 200 peer-reviewed research articles in *Nature Communications*, *Proceedings of the National Academy of Sciences of the United States of America*, *Journal of the American Chemical Society*, *Angewandte Chemie International Edition*, *ACS Nano*, *Nanoscale*, *The Journal of Physical Chemistry Letters*, and other international scientific research journals.

• • •