Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification

Yu Zhang¹, Zhihong Shen², Chieh-Han Wu², Boya Xie², Junheng Hao³, Ye-Yi Wang², Kuansan Wang², Jiawei Han¹

¹University of Illinois at Urbana-Champaign  ²Microsoft  ³University of California, Los Angeles
¹{yuz9, hanj}@illinois.edu,  ²{zhihosh, chiewu, boxie, yeyiwang, kuansanw}@microsoft.com,  ³jhao@cs.ucla.edu

ABSTRACT

Large-scale multi-label text classification (LMTC) aims to associate a document with its relevant labels from a large candidate set. Most existing LMTC approaches rely on massive human-annotated training data, which are often costly to obtain and suffer from a long-tailed label distribution (i.e., many labels occur only a few times in the training set). In this paper, we study LMTC under the zero-shot setting, which does not require any annotated documents with labels and only relies on label surface names and descriptions. To train a classifier that calculates the similarity score between a document and a label, we propose a novel metadata-induced contrastive learning (MICoL) method. Different from previous text-based contrastive learning techniques, MICoL exploits document metadata (e.g., authors, venues, and references of research papers), which are widely available on the Web, to derive similar document–document pairs. Experimental results on two large-scale datasets show that: (1) MICoL significantly outperforms strong zero-shot text classification and contrastive learning baselines; (2) MICoL is on par with the state-of-the-art supervised metadata-aware LMTC method trained on 10K–200K labeled documents; and (3) MICoL tends to predict more infrequent labels than supervised methods, thus alleviates the deteriorated performance on long-tailed labels.

CCS CONCEPTS

• Information systems → Data mining; • Computing methodologies → Classification and regression trees.

KEYWORDS

multi-label text classification, contrastive learning, metadata

ACM Reference Format:
Yu Zhang¹, Zhihong Shen², Chieh-Han Wu², Boya Xie², Junheng Hao³, Ye-Yi Wang², Kuansan Wang², Jiawei Han¹. 2022. Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. In Proceedings of the ACM Web Conference 2022 (WWW ’22), April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3485447.3512174

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW ’22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9096-5/22/04... $15.00
https://doi.org/10.1145/3485447.3512174

1 INTRODUCTION

Large-scale multi-label text classification (LMTC) [4] aims to find the most relevant labels to an input document given a large collection of candidate labels. As a fundamental task in text mining, LMTC has many Web-related applications such as product keyword recommendation on Amazon [6], academic paper classification on Microsoft Academic and PubMed [68], and article tagging on Wikipedia [18].

Most previous attempts address LMTC in a supervised fashion [1, 6, 17, 18, 22, 26, 32, 36, 40, 60, 62, 68], where the proposed text classifiers are trained on a large set of human-annotated documents. While achieving inspiring performance, these approaches have three limitations. First, obtaining enough human-labeled training data is often expensive and time-consuming, especially when the label space is large. Second, the trained classifiers can only predict labels they have seen in the training set. When new categories (e.g., “COVID-19”) emerge, the classifiers need to be re-trained. Third, the label distribution is often imbalanced in LMTC. Several labels (e.g., “World Wide Web”) have numerous training samples, while many others (e.g., “Bipartite Ranking”) occur only a few times. Related studies [54, 55] have shown that supervised approaches tend to predict frequent labels and overlook long-tailed ones.

Being aware of the annotation cost and frequent emergence of new labels, some studies [4, 16, 34, 38] focus on zero-shot LMTC. In their settings, annotated training documents are given for a set of seen classes, and they are tasked to build a classifier to predict

Figure 1: Two examples of labels with name(s) and description from Microsoft Academic [49] and PubMed [24].
unseen classes. However, as indicated in [61], we often face a more challenging scenario in real-world settings: all labels are unseen; we do not have any training samples with labels. For example, on Microsoft Academic [49], all labels are scientific concepts extracted from research publications on the Web [43], thus no annotated training data is available when the label space is created. Motivated by such applications, in this paper, we study zero-shot LMTC with fully new label space. The only signal we use to characterize each label is its name and description. Figure 1 shows two examples of label information from Microsoft Academic [49] and PubMed [24].

Although relevant labels are not available for documents, there is another type of information widely available on the Web but less concerned in previous studies: document metadata. We take scientific papers as an example: in addition to text information (e.g., title and abstract), a paper is also associated with various metadata fields such as its authors, venue, and references. As shown in Figure 2(a), a heterogeneous network [46] can be constructed to interconnect papers via their metadata information: papers, authors, and venues are nodes; authored-by, published-in, and cited-by are edges. Such metadata information could be strong label indicators of a paper. For example, in Figure 2(a), the venue node WWW suggests Doc1’s relevance to "World Wide Web". Moreover, metadata could also imply that two papers share common labels. For example, we know Doc1 and Doc2 may have similar research topics because they are co-authored by Andrew Tomkins and Ravi Kumar or they are co-cited by Doc3. More generally, metadata also exist in Web content such as e-commerce reviews (e.g., reviewer and product information) [64], social media posts (e.g., users and hashtags) [69], and code repositories (e.g., contributors) [70]. Although metadata have been used in fully supervised [68] or single-label [28, 64, 66, 67] text classification, it is largely unexplored in zero-shot LMTC.

Contributions. In this paper, we propose a novel metadata-induced contrastive learning (MICoL) framework for zero-shot LMTC. To perform classification, the key module in our framework is to compute a similarity score between two text units (i.e., one document and one label name/description) so that we can produce a rank list of relevant labels for each document. Without annotated document–label pairs to train the similarity scorer, we leverage metadata to generate similar document–document pairs. Inspired by the idea of contrastive learning [7], we train the scorer by pulling similar document closer while pushing dissimilar ones apart. For example, in Figure 2(a), we assume two papers sharing at least two authors are similar (this can be described by the notion of meta-paths [47] and meta-graphs [63], which will be formally introduced in Section 2.1). The similarity scorer is trained to score (Doc1, Doc2) higher than (Doc1, Doc8), where Doc8 is a randomly sampled paper. In the inference phase, as shown in Figure 2(b), we first use a discrete retriever (e.g., BM25 [39]) to select a set of candidate labels from the large label space. Next, we utilize the trained scorer to re-rank candidate labels to obtain the final classification results. Note that label information is only used during inference, thus no re-training is required when new labels emerge.

We demonstrate the effectiveness of MICoL on two datasets [68] extracted from Microsoft Academic [49] and PubMed [24], both with more than 15K labels. The results indicate that: (1) MICoL significantly outperforms strong zero-shot LMTC [61] and contrastive learning [8, 53, 57] baselines. (2) When we use P@k and NDCG@k as evaluation metrics, MICoL is competitive with the state-of-the-art supervised metadata-aware LMTC algorithm [68] trained on 10K–50K labeled documents; (3) When it is evaluated by metrics promoting correct prediction on tail labels [18, 55], MICoL is on par with the supervised method trained on 100K–200K labeled documents. This demonstrates that MICoL tends to predict more infrequent labels than supervised methods, thus alleviates the deteriorated performance on tail labels.

To summarize, this work makes the following contributions:

- We propose a zero-shot LMTC framework that utilizes document metadata. The framework does not require any labeled training data and only relies on label surface names and descriptions during inference.
- We propose a novel metadata-induced contrastive learning method. Different from previous contrastive learning approaches [14, 15, 25, 56, 58] which manipulate text only, we exploit metadata information to produce contrastive training pairs.
- We conduct extensive experiments on two large-scale datasets to demonstrate the effectiveness of the proposed MICoL framework.

2 PRELIMINARIES

2.1 Metadata, Meta-Path, and Meta-Graph

Metadata. Documents on the Web are usually accompanied by rich metadata information [28, 67, 68]. To provide a holistic view of documents with metadata, we can construct a heterogeneous information network (HIN) [46] to connect documents together.

Definition 2.1. (Heterogeneous Information Network [46]) An HIN is a graph $G = (V, E)$ with a node type mapping $\phi : V \rightarrow V_T$ and an edge type mapping $\psi : E \rightarrow E_T$. Either the number of node types $|V_T|$ or the number of edge types $|E_T|$ is larger than 1.

As shown in Figure 2(a), in our constructed HIN, each document is a node, and each metadata field is described by either a node (e.g., author, venue) or an edge (e.g., reference).
Reachability. We assume that two documents connected via a certain meta-path/meta-graph share similar topics. Formally, we introduce the concept of “reachable”.

**Definition 2.4.** (Reachable) Given a meta-path/meta-graph \( M \) and two documents \( d_1, d_2 \), we say \( d_2 \) is reachable from \( d_1 \) via \( M \) if and only if we can find a path/DAG \( M_0 \) in the HIN such that: (1) \( d_1 \) is the source node of \( M_0 \); (2) \( d_2 \) is the target node of \( M_0 \); (3) when each node in \( M_0 \) is abstracted by its node type, \( M_0 \) becomes \( M \).

We use \( d_1 \rightarrow_{M} d_2 \) to denote that \( d_2 \) is reachable from \( d_1 \) via \( M \), and we use \( N_M(d_1) \) to denote the set of nodes that are reachable from \( d_1 \) (i.e., \( N_M(d_1) = \{d_2 \mid d_1 \rightarrow_{M} d_2, d_2 \neq d_1\} \)).

### 2.2 Problem Definition

Zero-shot multi-label text classification [50] aims to tag each document with labels that are *unseen* during training time but available for prediction. Most previous studies [4, 16, 34, 38] assume that there is a set of *seen* classes, each of which has some annotated documents. Trained on these documents, their proposed text classifiers are expected to transfer the knowledge from seen classes to the prediction of unseen ones.

In this paper, we study a more challenging setting (proposed in [61] previously), where all labels are *unseen*. In other words, given the label space \( \mathcal{L} \), we do not have any training sample for \( l \in \mathcal{L} \). Instead, we assume a large-scale unlabeled corpus \( \mathcal{D} \) is given, and each document \( d \in \mathcal{D} \) is associated with metadata information. As mentioned in Section 2.1, with such metadata, we can construct an HIN \( G = (\mathcal{V}, \mathcal{E}) \) to describe the relations between documents. We aim to train a multi-label text classifier \( f \) based on both the text information \( \mathcal{D} \) and the network information \( G \). As an *inductive* task, the classifier \( f \) needs to predict its relevant labels given a new document \( d \notin \mathcal{D} \).

Since no training data is available to characterize a given label, same as previous studies [38, 61], we assume each label \( l \) has some text information to describe its semantics, such as label names \( n_l \) [42, 61] and descriptions \( s_l \) [3, 6].\(^2\) Examples of such label information have been shown in Figure 1. To summarize, our task can be formally defined as follows:

**Definition 2.5.** (Problem Definition) Given an unlabeled corpus \( \mathcal{D} \) with metadata information \( G = (\mathcal{V}, \mathcal{E}) \), and a label space \( \mathcal{L} \) with label names and descriptions \( (n_l, s_l) \in \mathcal{L} \), our task is to learn a multi-label text classifier \( f \) that can map a new document \( d \notin \mathcal{D} \) to its relevant labels \( \mathcal{D}_L \subseteq \mathcal{L} \).

### 3 THE MICOL FRAMEWORK

#### 3.1 A Two-Stage Framework

As shown in Figure 2(b), in the proposed MICOl framework, the LMTC problem is formulated as a ranking task. Specifically, given a new document (i.e., the “query”), our task is to predict top-ranked labels (i.e., the “items”) that are relevant to the document. Note that in LMTC, the label space \( \mathcal{L} \) (i.e., the “item pool”) is large. For example, in both the Microsoft Academic and PubMed datasets [68], there are more than 15,000 labels. Given a large item pool, recent ranking approaches are usually pipelined [13, 35], consisting of a first-stage discrete retriever (e.g., BM25 [39]) that efficiently generates a set

\(^{1}\)Following previous studies [45, 47], we view \( P_1 \rightarrow P_2 \rightarrow P_3 \) as a “directed path” from \( P_1 \) to \( P_3 \) by explaining it as \( P_1 \text{ cited by} P_2 \text{ cited by} P_3 \). In this way, \( P \rightarrow P \rightarrow P \) can be defined as a meta-path according to Definition 2.2. Similarly, we view \( \text{PAP} \) as a “directed path” \( P \text{ written by} A \text{ written by} P \). Using the same explanation, both \( \text{PAP} \) \( P \rightarrow \) \( P \) in Figure 3 can be viewed as a DAG, thus they are meta-graphs according to Definition 2.3.

\(^{2}\)Label names are required in our MICOl framework, but label descriptions are optional. That being said, MICOl is still applicable with label names only (i.e., \( s_l = \emptyset \)).
of candidate items followed by a continuous re-ranker (e.g., BERT [10]) that selects the most promising items from the candidates. Such design is a natural choice due to the effectiveness-efficiency trade-off among different ranking models: discrete rankers based on lexical matching are faster but less accurate; continuous rankers can perform latent semantic matching but are much slower.

Following such prevalent approaches, MICoL adopts a two-stage ranking framework, with a discrete retrieval stage and a continuous re-ranking stage. The major novelty of MICoL is that, with document metadata information, a new contrastive learning method is developed to significantly improve the re-ranking stage performance upon BERT-based models.

3.2 The Retrieval Stage

Since the main goal of this paper is to develop a novel contrastive learning framework for re-ranking, we do not aim at a complicated design of the retrieval stage. Therefore, we adopt two simple strategies: exact name matching and sparse retrieval.

Exact Name Matching. Given a document \(d\) and a label \(l\), if the label name \(n_l\) appears in the document text, we add \(l\) as a candidate of \(d\)'s relevant labels.\(^3\) We use \(C_{\text{exact}}(d)\) to denote the set of candidate labels obtained by exact name matching.

Sparse Retrieval. We cannot expect all relevant labels of a document explicitly appear in its text. To increase the recall of our retrieval stage, we adopt BM25 [39] to allow partial lexical matching between documents and labels. Specifically, we concatenate the name and the description together as the text information \(t_l\) of each label (i.e., \(t_l = n_l  \ |  \ s_l\)).\(^4\) Then, the score between \(d\) and \(l\) is calculated as

\[
BM25(d, l) = \sum_{w \in d \cap t_l} \text{IDF}(w) \cdot \frac{\text{TF}(w, t_l) \cdot (k_1 + 1)}{\text{TF}(w, t_l) \cdot k_1 \left(1 - b + b \frac{|d|}{\text{avgdl}}\right)}.
\]

Here, \(k_1 = 1.5\) and \(b = 0.75\) are parameters of BM25; \(\text{avgdl} = \frac{1}{|L|} \sum_{l \in L} |t_l|\) is the average length of label text information. Note that in classification tasks, documents are “queries” and labels are “items” being ranked. When the BM25 score between \(d\) and \(l\) exceeds a certain threshold \(\eta\), we add \(l\) as a candidate of \(d\)’s relevant labels. Formally,

\[
C_{\text{BM25}}(d) = \{l \mid l \in L, \ BM25(d, l) > \eta\}.
\]

Given a document \(d\), its candidate label set \(C(d) = C_{\text{exact}}(d) \cup C_{\text{BM25}}(d)\) (i.e., the union of candidates obtained by exact name matching and by sparse retrieval).

3.3 The Re-ranking Stage

Encouraged by the success of BERT [10] in a wide range of text mining tasks, we build our re-ranker upon BERT-based pre-trained language models. In general, our proposed re-ranking stage can be instantiated by any variant of BERT (e.g., SciBERT [2], BioBERT [23], and RoBERTa [21]). In our experiments, since documents from both Microsoft Academic and PubMed are scientific papers, we adopt SciBERT [2] as our building block.

\(^3\)On PubMed, as shown in Figure 1, each label can have more than one name because both "MeSH heading" and "entry terms" are viewed as label names. In this case, we add \(l\) as a candidate if any of its names appears in the document text.

\(^4\)On PubMed, instead of concatenating all label names into \(t_l\), we use the "MeSH heading" only as \(n_l\), which achieves better performance in experiments.

Figure 4: Two architectures that compute the similarity between a document \(d\) and a candidate label \(l\).

3.3.1 Bi-Encoder and Cross-Encoder. To improve the performance of BERT, two architectures are typically used for fine-tuning: Bi-Encoders and Cross-Encoders. Bi-Encoders [8, 37] perform self-attention over two text units (e.g., query and item) separately and compute the similarity between their representation vectors at the end. Cross-Encoders [13, 61], in contrast, perform self-attention within as well as across two text units at the same time. Below we introduce how to apply these two architectures to our task.

Bi-Encoder. Given a document \(d\) and a candidate label \(l \in C(d)\) (obtained in the retrieval stage), we use BERT to encode them separately to generate two representation vectors.

\[
e_d = \text{BERT}(d), \quad e_l = \text{BERT}(t_l).
\]

To be specific, given the document text \(d\) (resp., the label name and description \(t_l\)), we use the sequence "[CLS] \(d\) [SEP]" (resp., "[CLS] \(t_l\) [SEP]"") as the input into BERT and take the output vector of the "[CLS]" token from the last layer as the document representation \(e_d\) (resp., label representation \(e_l\)). The score between \(d\) and \(l\) is defined as the cosine similarity of their representation vectors.

\[
\text{score}(d, l) = \cos(e_d, e_l).
\]

Cross-Encoder. To better utilize the fully connected attention mechanism of BERT-based models, we can concatenate document and label text information together and encode it using one BERT.

\[
e_{d|l} = \text{BERT}(d \ | \ t_l).
\]

Here, \((d \ | \ t_l)\) denotes the input sequence "[CLS] \(d\) [SEP] \(t_l\) [SEP]". Again, we take the output vector of the "[CLS]" token as \(e_{d|l}\). The score between \(d\) and \(l\) is then obtained by adding a linear layer upon BERT.

\[
\text{score}(d, l) = \mathbf{w}^\top \mathbf{e}_{d|l},
\]

where \(\mathbf{w}\) is a trainable vector.

The architectures of Bi-Encoder and Cross-Encoder are illustrated in Figure 4.

3.3.2 Metadata-Induced Contrastive Learning (MICoL). Now we aim to fine-tune Bi-Encoder and Cross-Encoder to improve their re-ranking performance. (For Cross-Encoder, especially, we cannot even run it without fine-tuning because \(\mathbf{w}\) needs to be learned.) If our task were fully supervised, we would have positive document–label training pairs \((d, l)\) indicating \(d\) is labeled with \(l\), and the training objective would be maximizing \(\text{score}(d, l)\) for these positive pairs. However, we do not have any annotated documents with the zero-shot setting. In this case, to fine-tune above two architectures, we adopt a contrastive learning framework.
Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification

WW‘22, April 25–29, 2022, Virtual Event, Lyon, France

Instead of learning “what is what”, contrastive learning [7] tries to learn “what is similar to what”. In our problem setting, we assume there is a collection of document pairs \((d, d^+)\), where \(d\) and \(d^+\) are similar to each other, e.g., \(d^+\) is reachable from \(d\) via a specified meta-path or meta-graph. For each \(d\), we can also randomly sample a set of documents \(\{d_i^+\}_{i=1}^N\) from the whole corpus \(D\). Contrastive learning aims to learn effective representations by pulling \(d\) and \(d^+\) together while pushing \(d\) and \(d_i^-\) apart. Taking Bi-Encoder as an example, we first use BERT to encode all documents.

\[
e_d = \text{BERT}(d), \quad e_{d^+} = \text{BERT}(d^+), \quad e_{d_i^-} = \text{BERT}(d_i^-).
\] (7)

Following Chen et al.’s seminal work [7], the contrastive loss can be defined as

\[
\begin{equation}
\mathcal{J}_\text{mi} = \mathbb{E}_{d^+ \in \mathcal{N}_M(d)} \left[ -\log \frac{\exp(\cos(e_d, e_{d^+})/\tau)}{\sum_{d' \in \mathcal{N}_M(d)} \exp(\cos(e_d, e_{d'})/\tau)} \right].
\end{equation}
\] (9)

Similarly, for Cross-Encoder, we first compute \(\text{score}(d, d^+\rangle\) and \(\text{score}(d, d_i^-\rangle\).

\[
\mathcal{J}_\text{cross} = \mathbb{E}_{d^+ \in \mathcal{N}_M(d)} \left[ -\log \frac{\exp(\text{score}(d, d^+\rangle)}{\sum_{d' \in \mathcal{N}_M(d)} \exp(\text{score}(d, d_i^-\rangle)} \right].
\] (11)

The BERT model is thus fine-tuned by minimizing the contrastive loss in Eq. (9) or (11). Figure 5 illustrates the fine-tuning process of both Bi-Encoder and Cross-Encoder using \(M = P(\text{AA})P\).

Due to the space constraint, the training and inference procedures of the MICO framework are formally summarized in Appendix A.2; the optimization details of MICO (i.e., Eqs. (9) and (11)) are provided in Appendix A.3.

4 EXPERIMENTS

4.1 Setup

Datasets. Given the task of metadata-aware LMTC, following [68], we perform evaluation on two large-scale datasets.

- **MAG-CS [49]**. The Microsoft Academic Graph (MAG) has a web-scale collection of scientific papers from various fields. In [68], 705,407 MAG papers published at 105 top CS conferences from 1990 to 2020 are selected to form a dataset with 15,808 labels.

- **PubMed [24]**. PubMed has a web-scale collection of biomedical literature from MEDLINE, life science journals, and online books. In [68], 898,546 PubMed papers published in 150 top medicine journals from 2010 to 2020 are selected to form a dataset with 17,963 labels (i.e., MeSH terms [9]).

Under the fully supervised setting, Zhang et al. [68] split both datasets into training, validation, and testing sets. In this paper, we focus on the zero-shot setting. Therefore, we combine their training and validation sets together as our unlabeled input corpus \(D\) (that being said, we do not know the labels of these documents, and we only utilize their text and metadata information). We use their testing set as our testing documents \(d \notin D\). Dataset statistics are briefly listed in Table 1. More details are in Appendix A.4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Training (Unlabeled)</th>
<th>#Testing</th>
<th>#Labels</th>
<th>Labels/Doc</th>
<th>Words/Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAG-CS</td>
<td>634,874</td>
<td>78,533</td>
<td>15,808</td>
<td>5.59</td>
<td>126.55</td>
</tr>
<tr>
<td>PubMed</td>
<td>808,692</td>
<td>89,854</td>
<td>17,963</td>
<td>7.80</td>
<td>199.14</td>
</tr>
</tbody>
</table>

**Compared Methods.** We evaluate MICO against a variety of baseline methods using text embedding, pre-trained language models, and text-based contrastive learning. Since the major technical contribution of MICO is in the re-ranking stage, all the baselines

\[1\] Originally, there were 15,809 labels in MAG-CS, but the label "Computer Science" is removed from all papers because it is trivial to predict.
Table 2: P@$k$ and NDCG@$k$ scores of compared algorithms on MAG-CS and PubMed. Bold: the highest score of zero-shot approaches. *: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.05. **: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.01.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAG-CS [40]</th>
<th>PubMed [24]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@1</td>
<td>P@3</td>
</tr>
<tr>
<td>Doc2Vec [20]</td>
<td>0.6907**</td>
<td>0.4613**</td>
</tr>
<tr>
<td>SciBERT [2]</td>
<td>0.6440**</td>
<td>0.5030**</td>
</tr>
<tr>
<td>ZeroShot-Entail [61]</td>
<td>0.7077**</td>
<td>0.5381**</td>
</tr>
<tr>
<td>SPECTER [8]</td>
<td>0.6442**</td>
<td>0.4939**</td>
</tr>
<tr>
<td>EDA [53]</td>
<td>0.6291**</td>
<td>0.4848**</td>
</tr>
<tr>
<td>UDA [97]</td>
<td>0.5965**</td>
<td>0.3950**</td>
</tr>
<tr>
<td>ZeroShot (Bi-Encoder, $P \rightarrow P$)</td>
<td>0.7062**</td>
<td>0.5369**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, $P \rightarrow (PP)$ → $P$)</td>
<td>0.7795**</td>
<td>0.5344**</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, $P \rightarrow (PP)$ → $P$)</td>
<td>0.7177**</td>
<td>0.5444**</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)</td>
<td>0.7061**</td>
<td>0.5376**</td>
</tr>
</tbody>
</table>

We also report the performance of a fully supervised method for reference.

- **MATCH [68]** is the state-of-the-art supervised approach for metadata-aware multi-label text classification. Because we do not consider label hierarchy in our problem setting, we report the performance of MATCH-NoHierarchy with various sizes of training data for comparison.

**Evaluation Metrics.** Following the commonly used evaluation on multi-label text classification [22, 62, 68], we adopt two rank-based metrics: P@$k$ and NDCG@$k$, where $k = 1, 3, 5$. For a document $d$, let $y_d \in \{0, 1\}^{|L|}$ be its ground truth label vector and rank$(i)$ be the index of the $i$-th highest predicted label according to the re-ranker.

$$P@k = \frac{1}{k} \sum_{i=1}^{k} y_{d,\text{rank}(i)}.$$  

$$\text{NDCG}@k = \frac{\sum_{i=1}^{k} \frac{\log(1+i)}{\log(1+i)}}{\sum_{i=1}^{k} \frac{\log(1+i)}{\log(1+i)}}.$$  

**4.2 Performance Comparison**

Table 2 shows P@$k$ and NDCG@$k$ scores of compared algorithms on MAG-CS and PubMed. We run each experiment three times with the average score reported. (SciBERT, ZeroShot-Entail, and SPECTER are deterministic according to our usage, so we run them only once.) To show statistical significance, we conduct two-tailed unpaired t-tests to compare the best performed MICoL model and other approaches including MATCH. (When comparing MICoL with three deterministic approaches, we conduct two-tailed Z-tests instead.) The significance level of each result is marked in Table 2. For MICoL, we show the performance of one meta-path $P \rightarrow P \leftarrow P$ and one meta-graph $P \leftarrow (PP) \rightarrow P$ here. The performance of other meta-paths/meta-graphs will be presented in Table 4 and discussed in Section 4.4.

From Table 2, we observe that: (1) MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) significantly outperforms all zero-shot baselines in most cases, except that it is slightly worse than ZeroShot-Entail on PubMed in terms of P@5 and NDCG@5. (2) On MAG-CS, MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) performs significantly better than the supervised MATCH model with 50k labeled training data.
Table 3: PSP@k and PSN@k scores of compared algorithms on MAG-CS and PubMed. Bold, *; and **: the same meaning as in Table 2. We also show the ratio PSP@1/P@1. The higher PSP@k/P@k is, the more infrequent the correctly predicted labels are.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@1</td>
<td>P@3</td>
</tr>
<tr>
<td>DocVec [31]</td>
<td>0.4357**</td>
<td>0.4623**</td>
</tr>
<tr>
<td>SciBERT [2]</td>
<td>0.4688**</td>
<td>0.4958**</td>
</tr>
<tr>
<td>ZeroShot-Entail [61]</td>
<td>0.4976**</td>
<td>0.4892**</td>
</tr>
<tr>
<td>SPECITER [8]</td>
<td>0.5084</td>
<td>0.5334*</td>
</tr>
<tr>
<td>EDA [51]</td>
<td>0.4951**</td>
<td>0.4968**</td>
</tr>
<tr>
<td>UDIA [57]</td>
<td>0.4850**</td>
<td>0.4907**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P ← P)</td>
<td>0.5176</td>
<td>0.5311</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P ← (PP) → P)</td>
<td>0.5160</td>
<td>0.5281</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → P ← P)</td>
<td>0.5375</td>
<td>0.5415</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → (PP) → P)</td>
<td>0.5326</td>
<td>0.5363</td>
</tr>
</tbody>
</table>

On PubMed, MICoL can be competitive with MATCH trained on more than 10k annotated documents in terms of P@3, P@5, and NDCG@5. (3) Using purely text-based augmentation approaches (i.e., EDA and UDIA) to perform contrastive learning is not consistently beneficial. In fact, EDA and UDIA perform even worse than unfine-tuned SciBERT on MAG-CS. In contrast, our proposed metadata-induced contrastive learning consistently boosts the performance of SciBERT, and the improvements are much more significant than those of text-based contrastive learning. (4) For both P → P ← P and P ← (PP) → P, Cross-Encoder performs better than Bi-Encoder within the MICoL framework. In Section 4.4, we will show that this observation is generalizable to most meta-paths and meta-graphs we use.

4.3 Performance on Tail Labels

Tail labels refer to those labels relevant to only a few documents in the dataset. They are usually more fine-grained and informative than head labels (i.e., frequent ones). However, predicting tail labels is less “rewarding” for models to achieve high P@k and NDCG@k scores. Therefore, new scoring functions are designed to promote infrequent label prediction by giving the model a higher reward when it predicts a tail label correctly. Propensity-scored P@k (PSDCG@k) and propensity-scored NDCG@k (PSNDCG@k, abbreviated to PSN@k in this paper) are thus proposed in [18] and widely used in LMTc evaluation [16, 32, 40, 55, 62]. P@k and PSN@k are defined as follows.6

\[
\frac{1}{P_i} = 1 + C (N_j + B)^{-A}, \quad P@k = \frac{1}{k} \sum_{i=1}^{k} \frac{y_{d\text{rank}(i)}}{p_{d\text{rank}(i)}}
\]

\[
PSDCG@k = \frac{1}{k} \sum_{i=1}^{k} \frac{y_{d\text{rank}(i)}}{p_{d\text{rank}(i)} \log(i + 1)}, \quad PSN@k = \frac{PSDCG@k}{\sum_{i=1}^{k} \frac{y_{d\text{rank}(i)}}{\log(i + 1)}}
\]

Here, \( \frac{1}{P_i} \) is the “reward” of predicting the label \( l \) correctly; \( N_j \) is the number of documents relevant to \( l \) in the training set. Following previously established parameter values [18, 55, 62], we set \( A = 0.55, B = 1.5, C = (\log|D| - 1)(B + 1)^{-A} \). Therefore, the less frequent a label is, the higher reward one can get by predicting it correctly. Table 3 shows the P@k and PSN@k scores of compared methods. As shown in Table 3, when we use PS@k and PSN@k as evaluation metrics, MICoL becomes more powerful. MICoL (Cross-Encoder, P → P ← P) significantly outperforms all zero-shot baselines, and it is on par with MATCH trained on 100k–200k labeled documents. According to the definition of PS@k and P@k, the ratio PS@k/P@k reflects the average “reward” a model gets from its correctly predicted labels. The higher PS@k/P@k is, the more infrequent the correctly predicted labels are. We show PS@1/P@1 in Table 3. We observe that labels predicted by MICoL (and all other zero-shot methods) are much more infrequent than labels predicted by the supervised MATCH model. The reason could be that zero-shot methods cannot see any labeled data during training, thus they get no hints of frequent labels and are not biased towards head labels. This helps alleviate the deteriorated performance of supervised models on long-tailed labels as observed in [54, 55].

4.4 Effect of Meta-Path and Meta-Graph

Table 4 shows the performance of all 20 MICoL variants (2 architectures × 10 meta-paths/meta-graphs). We have the following observations: (1) All meta-paths and meta-graphs used in MICoL, except PVP, can improve the classification performance upon unfine-tuned SciBERT. For PVP, the unsatisfying performance is expected because venue information alone (e.g., ACL) is too weak to distinguish between fine-grained labels (e.g., “Named Entity Recognition” and “Entity Linking”). In Appendix A.1, we provide a mathematical interpretation of why some meta-paths/meta-graphs (e.g., PVP or P(AAAAA)P) may not perform well within our MICoL framework. In short, the results in Table 4 demonstrate the effectiveness of MICoL across different meta-paths/meta-graphs. (2) Cross-Encoder models perform better than their Bi-Encoder counterparts in most cases (8 out of 10 meta-paths/meta-graphs on MAG-CS and 10 out of 10 on PubMed, in terms of P@1). (3) In contrast to the gap between Bi-Encoder and Cross-Encoder, the difference among citation-based meta-paths and meta-graphs is less significant. It would be an interesting future work to automatically select the most effective meta-paths/meta-graphs, although related studies on heterogeneous network representation learning [11, 51, 59] often require users to specify them.
Table 4: P@k and NDCG@k scores of MICoL using different meta-paths/meta-graphs. Bold: the best model. *: significantly worse than the best model with p-value < 0.05. **: significantly worse than the best model with p-value < 0.01. All meta-paths and meta-graphs, except PVP, can improve the classification performance upon unfine-tuned SciBERT.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P@1</th>
<th>P@3</th>
<th>P@5</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfine-tuned SciBERT</td>
<td>0.6399*</td>
<td>0.5113*</td>
<td>0.4566*</td>
<td>0.5610*</td>
<td>0.5136*</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, PAP)</td>
<td>0.6877**</td>
<td>0.5285**</td>
<td>0.4143**</td>
<td>0.5852**</td>
<td>0.5280**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, PVP)</td>
<td>0.6589**</td>
<td>0.5123**</td>
<td>0.4063**</td>
<td>0.5656**</td>
<td>0.5145**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P)</td>
<td>0.7094*</td>
<td>0.5391*</td>
<td>0.4190*</td>
<td>0.5982*</td>
<td>0.5367*</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P)</td>
<td>0.7062*</td>
<td>0.5360*</td>
<td>0.4184*</td>
<td>0.5960*</td>
<td>0.5355*</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P)</td>
<td>0.7039**</td>
<td>0.5379**</td>
<td>0.4187**</td>
<td>0.5963*</td>
<td>0.5356*</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P)</td>
<td>0.6873**</td>
<td>0.5272**</td>
<td>0.4130**</td>
<td>0.5840**</td>
<td>0.5269**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → P)</td>
<td>0.6832**</td>
<td>0.5263**</td>
<td>0.4135**</td>
<td>0.5823**</td>
<td>0.5263**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → (PP) → P)</td>
<td>0.7015**</td>
<td>0.5334**</td>
<td>0.4160**</td>
<td>0.5928*</td>
<td>0.5322**</td>
</tr>
<tr>
<td>MICoL (Bi-Encoder, P → (PP) → P)</td>
<td>0.7050**</td>
<td>0.5344**</td>
<td>0.4161†</td>
<td>0.5932*</td>
<td>0.5331†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, PAP)</td>
<td>0.7034†</td>
<td>0.5355†</td>
<td>0.4168†</td>
<td>0.5943†</td>
<td>0.5337†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, PVP)</td>
<td>0.6720*</td>
<td>0.5203*</td>
<td>0.4103*</td>
<td>0.5750†</td>
<td>0.5210†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → P)</td>
<td>0.7033†</td>
<td>0.5391†</td>
<td>0.4201†</td>
<td>0.5974†</td>
<td>0.5365†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → P)</td>
<td>0.7169*</td>
<td>0.5409*</td>
<td>0.4214†</td>
<td>0.6033*</td>
<td>0.5406*</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → P)</td>
<td>0.7177†</td>
<td>0.5444†</td>
<td>0.4219†</td>
<td>0.6048†</td>
<td>0.5415†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → P)</td>
<td>0.7045†</td>
<td>0.5356†</td>
<td>0.4168†</td>
<td>0.5944†</td>
<td>0.5336†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → (PP) → P)</td>
<td>0.7028†</td>
<td>0.5351†</td>
<td>0.4171†</td>
<td>0.5939†</td>
<td>0.5338†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → (PP) → P)</td>
<td>0.7024†</td>
<td>0.5354†</td>
<td>0.4177†</td>
<td>0.5940*</td>
<td>0.5343†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → (PP) → P)</td>
<td>0.7076†</td>
<td>0.5379†</td>
<td>0.4188†</td>
<td>0.5971†</td>
<td>0.5363†</td>
</tr>
<tr>
<td>MICoL (Cross-Encoder, P → (PP) → P)</td>
<td>0.7061†</td>
<td>0.5376†</td>
<td>0.4187†</td>
<td>0.5964†</td>
<td>0.5357†</td>
</tr>
</tbody>
</table>

5 RELATED WORK

Zero-Shot Multi-Label Text Classification. Yin et al. [61] divide existing studies on zero-shot text classification into two settings: the restrictive setting [50] assumes training documents are given for some seen classes and the trained classifier should be able to predict unseen classes; the wild setting does not assume any seen classes and the classifier needs to make prediction without making any annotated training data. Most previous studies on zero-shot LMTC [4, 16, 33, 34, 38] focus on the restrictive setting, while this paper studies the more challenging wild setting. Under the wild setting, a pioneering approach is dataless classification [5, 44] which maps documents and labels into the same space of Wikipedia concepts. Recent studies further leverage convolutional networks [29] or pre-trained language models [27, 30, 52]. While these models rely on label names only, they assume that each document belongs to one category, thus are not applicable to the multi-label setting. Yin et al. [61] propose to treat zero-shot text classification as a textual entailment problem and leverage pre-trained language models to solve it; Shen et al. [42] further extend the idea to hierarchical zero-shot text classification. Compared with their studies, we utilize document metadata as complementary signals to the text.

Metada-Aware Text Classification. In some specific classification tasks, document metadata have been used as label indicators (e.g., reviewer and product information in review sentiment analysis [48], user profile information in tweet localization [41, 69]). Kim et al. [19] propose a generic approach to add categorical metadata into neural text classifiers. Zhang et al. [68] further study metadata-aware LMTC. However, all these studies focus on the fully supervised setting. Some studies [64–67] leverage metadata in few-shot text classification. Nevertheless, in LMTC, since the label space is large, it becomes prohibitive to provide even a few training samples for each label. Mekala et al. [28] consider metadata in zero-shot single-label text classification given a small label space, but their approach can hardly be generalized to LMTC.

Text Contrastive Learning. Recently, contrastive learning has become a promising trend in unsupervised text representation learning. The general idea is to use various data augmentation techniques [12, 53, 57] to generate similar text pairs and find a mapping function to make them closer in the representation space while pushing away dissimilar ones. Therefore, the major novelty of those studies is often the data augmentation techniques they propose. For example, DeCLUTR [15] samples two text spans from the same document; CLEAR [56] adopts span deletion, span reordering, and synonym substitution; DECA [25] uses synonym substitution, antonym augmentation, and back translation; SimCSE [14] applies two different hidden dropout masks when encoding the same sentence; ConSERT [58] employs token reordering, deletion, dropout, and adversarial attack. However, all these data augmentation approaches manipulate text information only to generate artificial sentences/documents, while our MICoL exploits metadata information to find real documents that are similar to each other.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we study zero-shot multi-label text classification with document metadata as complementary signals, which allow us to utilize heterogeneous network information besides corpus. Our setting does not require any annotated training documents with labels and only relies on labels names and descriptions. We propose a novel metadata-induced contrastive learning (MICoL) method to train a BERT-based document–label relevance scorer. We study two types of architectures (i.e., Bi-Encoder and Cross-Encoder) and 10 different meta-paths/meta-graphs within the MICoL framework. MICoL achieves strong performance on two large datasets, outperforming competitive baselines under the zero-shot setting and being outperformed by our supervised MATCH model trained on 10K–200K labeled documents. In the future, it is of interest to extend MICoL to zero-shot hierarchical text classification and explore whether label hierarchy could provide additional signals to contrastive learning.
ACKNOWLEDGMENTS

We thank anonymous reviewers for their valuable and insightful feedback. Research was supported in part by US DARPA KAROS Program No. FA8750-19-2-1004, SocialSim Program No. W911NF-17-C-0069, and INCAS Program No. HR00121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS-17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 1908997. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES


[9] Margaret H Coletti and Howard L Bleich. 2001. Medical subject headings used to express or implied, of DARPA or the U.S. Government.


[31] Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. 2011. Distributed representations of words and phrases and their compositionality. In EMNLP’11. 830–835.


[59] Carl Yang, Yuxin Xiao, Yu Zhang, Yihsiu Sun, and Jiawei Han. 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. IEEE TKDE (2020).
[67] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F. Xu, Xuan Wang, and Jiawei Han. 2020. Minimally Supervised Categorization of Text with Metadata. In SIGIR’20. 1231–1240.
[68] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In WWW’21. 3246–3257.
[70] Yu Zhang, Frank F. Xu, Sha Li, Yu Meng, Xuan Wang, Qi Li, and Jiawei Han. 2019. HiGitClass: Keyword-Driven Hierarchical Classification of GitHub Repositories. In ICDM’19. 876–885.
A  SUPPLEMENTARY MATERIAL
A.1  Interpretation of MICoL

We have proposed the MICoL objective (i.e., Eq. (9) or (11)) to fine-tune BERT. Here, we provide the interpretation of this objective function by answering the following two questions.

• Q1. MICoL does not require any labeled data during fine-tuning. What is the relationship between MICoL and supervised fine-tuning approaches?

• Q2. Under what conditions will the MICoL objective be closer to the supervised objective (thus MICoL may perform better in classification)?

To answer these two questions, let us first consider how supervised approaches fine-tune BERT in multi-label text classification. Assume each document \(d \in D\) is annotated with its relevant labels \(L_d\). We can sample a positive label \(l^*\) from \(L_d\) and several negative labels \(l_i^*\) \(\forall i \neq 1\) from \(L\). The supervised learning process learns effective representations by pulling \(d\) and \(l^*\) together while pushing \(d\) and \(l_i^*\) apart. Using either Bi-Encoder (Eqs. (3) and (4)) or Cross-Encoder (Eqs. (5) and (6)), the learning objective can be defined as

\[
J_{Sup} = \mathbb{E}_{i \in [1,M]} \left[ -\log \frac{\exp(score(d,l^*))}{\exp(score(d,l^*)) + \sum_{i \neq 1} \exp(score(d,l_i^*))} \right]. \tag{12}
\]

If we view each label \(l\)’s text information (i.e., label name and description) as a “document”, it is natural to assume this “document” \(t_l\) is relevant to the label \(l\). For example, in Figure 1, the description of “Webgraph” can be viewed as a “document” specifically relevant to “Webgraph” itself. Therefore, if we follow the notation of \(L_d\) and use \(L_l\) to denote the set of labels relevant to \(t_l\), it is equivalent to have \(L_l = \{l\}\). In this case, we have \(l \in L_d\) if and only if \(L_d \cap L_l \neq \emptyset\). Then, the supervised loss in Eq. (12) is equivalent to

\[
J_{Sup} = \mathbb{E}_{i \in \mathbb{L}_d \setminus L_l \neq \emptyset} \left[ -\log \frac{\exp(score(d,l^*))}{\exp(score(d,l^*)) + \sum_{i \neq 1} \exp(score(d,l_i^*))} \right]. \tag{13}
\]

If we replace \(l^*\) and \(l_i^*\) in Eq. (13) with \(d^*\) and \(d_i^*\), respectively, (in other words, if we replace label text information with real documents), the supervised loss can be rewritten into the following form.

\[
J_{Sup}' = \mathbb{E}_{d_i^* \in D \setminus L_d \setminus L_l \neq \emptyset} \left[ -\log \frac{\exp(score(d^*,d^*))}{\exp(score(d^*,d^*)) + \sum_{i \neq 1} \exp(score(d^*,d_i^*))} \right]. \tag{14}
\]

By comparing Eq. (14) with Eq. (9) when \(\tau = 1\) and Eq. (11), we can answer Q1. MICoL and supervised loss share the same functional format, and the only difference is the distribution of training data \((d,d^*)\) pairs. For supervised loss, annotated documents are available, so positive document–document pairs \((d,d^*)\) can be derived using their labels; for MICoL, there are no annotated training samples, positive document–document pairs \((d,d^*)\) are derived from metadata instead.

Now we proceed to Q2. Given the answer of Q1, the key difference between \(J_{Bi}\) (or \(J_{Cross}\)) and \(J_{Sup}'\) is how we sample the positive partner \(d^*\) for each document \(d\). Let us explicitly write down the two distributions for sampling \(d^*\). For MICoL, the distribution is

\[
P_{MICoL}(d^*|d) = \begin{cases} 1/X, & d^* \in N_M(d); \\ 0, & d^* \notin N_M(d). \end{cases} \tag{15}
\]

Here, \(X = |N_M(d)|\).

For the supervised loss in Eq. (14), the distribution is

\[
P_{Sup}(d^*|d) = \begin{cases} 1/Y, & L_d \cap L_{d^*} \neq \emptyset; \\ 0, & L_d \cap L_{d^*} = \emptyset. \end{cases} \tag{16}
\]

Here, \(Y = \{|d^*| \cap L_d \neq \emptyset\}\).

It is straightforward to compute the Jensen-Shannon (JS) divergence between \(P_{MICoL}\) and \(P_{Sup}\):

\[
JS(P_{MICoL}||P_{Sup}) = \frac{1}{2} \log \frac{2X}{X+Y} + \frac{1}{2} \sum_{d^* \in N_M(d)} \frac{1}{X} \log \left( 1 + \frac{X}{Y} \right) + \frac{1}{2} \sum_{d^* \notin N_M(d)} \frac{1}{Y} \log \left( 1 + \frac{Y}{X} \right). \tag{17}
\]

When \(JS(P_{MICoL}||P_{Sup})\) is small, \(P_{MICoL}\) is close to \(P_{Sup}\), thus the MICoL objective \(J_{Bi}\) or \(J_{Cross}\) is similar to the supervised objective \(J_{Sup}\). To make \(JS(P_{MICoL}||P_{Sup})\) smaller, the meta-path/meta-graph \(M\) should satisfy the following two properties as close as possible:

**Property 1:** \(d^* \in N_M(d) \Rightarrow L_d \cap L_{d^*} \neq \emptyset\). Intuitively, this property means that if \(d^*\) is reachable from \(d\) via \(M\), then \(d\) and \(d^*\) should have similar topic labels. When this property is perfectly satisfied, the second term in Eq. (17) is 0.

**Property 2 (the inverse of Property 1):** \(L_d \cap L_{d^*} \neq \emptyset \Rightarrow d^* \in N_M(d)\). This property implies that if \(d\) and \(d^*\) have similar labels, then \(d^*\) should be reachable from \(d\) via \(M\). When this property is perfectly satisfied, the fourth term in Eq. (17) is 0.

For example, when the label space is fine-grained, using \(M = P\) may not be a good choice because sharing the same venue may not be suitable because two papers in the same category do not necessarily have so many common authors. In this case, \(P(AAAAA)P\) may not be suitable because two papers in the same category do not necessarily have so many common authors. In this case, \(P(AAAAA)P\) violates Property 2.

A.2  Training and Inference Procedures of MICoL

We summarize the complete training and inference procedures in Algorithms 1 and 2, respectively. The goal of training is to obtain a fine-tuned BERT model, which is later used in the re-ranking stage of inference. Note that during the model training phase, we calculate the similarity score between two documents, while during the inference phase, we calculate the score between a document and a label (represented by its name and description).

A.3  Implementation

A.3.1  Optimization of MICoL

To approximate the expectation in Eqs. (9) and (11) during optimization, we adopt a sampling strategy. Specifically, we first sample a set of documents from \(D\). For each
A.3.2 Parameter Settings of MICoL. For both Bi-Encoder and Cross-Encoder, we sample 50,000 \((d, d^*)\) pairs for training and 5,000 \((d, d^*)\) pairs for validation. The training batch size \(\beta\) is 8 and 4 for Bi-Encoder and Cross-Encoder, respectively. The maximum length of SciBERT in Bi-Encoder is 256; the maximum length of SciBERT in Cross-Encoder is 512 (i.e., 256 for each document before concatenation). The temperature hyperparameter \(\tau = 0.05\). We train the model for 3 epochs using Adam as the optimizer. During inference, the threshold of BM25 scores is \(\eta = 400\).

A.4 Datasets

Table 1 mainly summarizes text and label statistics (where Labels/Doc and Words/Doc refer to the statistics in the testing set). In Table 5, we list metadata-related statistics of the training corpus \(\mathcal{D}\).

Table 5: Metadata-related statistics of the training corpus.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Authors</td>
<td>762,259</td>
<td>2,068,411</td>
</tr>
<tr>
<td># Author–Paper Edges</td>
<td>2,047,166</td>
<td>5,391,314</td>
</tr>
<tr>
<td># Venues</td>
<td>105</td>
<td>159</td>
</tr>
<tr>
<td># Venue–Paper Edges</td>
<td>634,874</td>
<td>808,692</td>
</tr>
<tr>
<td># Paper→Paper Edges</td>
<td>1,219,234</td>
<td>3,615,220</td>
</tr>
</tbody>
</table>

A.5 Additional Experiments: Combining Top-Performing Meta-Paths/Meta-Graphs

Figure 6 compares P@k and NDCG@k scores of top-3 performing meta-paths/meta-graphs and the combination of them. We combine them by putting document–document pairs induced by different meta-paths/meta-graphs together to train the model. In Figure 6, we cannot observe consistent and significant benefit of combining meta-paths/meta-graphs, possibly because two documents judged as similar by one meta-path may be viewed as dissimilar by another, which may confuse our model during training.

```
Algorithm 1: MICoL Training
Input: An unlabeled corpus \(\mathcal{D}\) with metadata \(G = (\mathcal{V}, \mathcal{E})\); a meta-path/meta-graph \(M\); the pre-trained BERT model.
Output: A fine-tuned BERT model.
1 Sample \(d, d^*\), and \(d_i^*\) from \(\mathcal{D}\) according to \(M\);
2 if using Bi-Encoder for fine-tuning then
   \(e_d, e_{d^*}, \{e_{d_i^*}\}_{i=1}^{N}\) ← Eq. (7);
3 Fine-tuning BERT by optimizing Eq. (9);
4 else if using Cross-Encoder for fine-tuning then
   \(\text{score}(d, d^*), \text{score}(d, d_i^*)\) ← Eq. (10);
5 Fine-tuning BERT by optimizing Eq. (11);
6 Return the fine-tuned BERT;
```

Algorithm 2: MICoL Inference
Input: A label space \(\mathcal{L}\) with label names and descriptions \(t_l\); the fine-tuned BERT model; a new document \(d \notin \mathcal{D}\).
Output: \(d\)’s relevant labels \(L_d \subseteq \mathcal{L}\).
1 // The Retrieval Stage;
2 \(C_{\text{exact}}(d)\) ← exact label name matching;
3 \(C_{\text{BM25}}(d)\) ← Eq. (2);
4 \(C(d) = C_{\text{exact}}(d) \cup C_{\text{BM25}}(d)\);
5 // The Re-ranking Stage;
6 for \(l \in C(d)\) do
   7 if using Bi-Encoder then
      \(e_d, e_l\) ← Eq. (3) using the fine-tuned BERT model;
   8 \(\text{score}(d, l)\) ← Eq. (4);
   10 else if using Cross-Encoder then
      \(e_{d,l}\) ← Eq. (5) using the fine-tuned BERT model;
   11 \(\text{score}(d, l)\) ← Eq. (6);
12 Rank \(l \in C(d)\) according to \(\text{score}(d, l)\);
13 Return \(L_d = \{\text{top}-k \text{ ranked labels in } C(d)\}\);
```

sampled document \(d\), we pick one \(d^*\) from \(N_M(d)\) as \(d^*\)’s positive partner (if \(N_M(d) = \emptyset\)). When fine-tuning the Bi-Encoder, following [7], \(d^*\)’s negative partners \(d_i^*\) are the positive partner of other documents in the same training batch. Each \(d\) has 1 positive partner and \((\beta - 1)\) negative partners, where \(\beta\) is the batch size. Therefore, within each batch, to calculate Eq. (9), we need to call BERT \(2\beta\) times (i.e., computing \(e_d\) and \(e_{d_i^*}\) for each \(d\)). When fine-tuning the Cross-Encoder, however, we would call BERT \(\beta^2\) times within each batch if we used the same approach to obtain negative partners. To make the optimization process more efficient, for each \(d\), we directly sample one negative partner \(d^*\) from \(\mathcal{D}\). Each \(d\) has 1 positive partner and 1 negative partner. In this way, we only need to call BERT \(2\beta\) times within each batch (i.e., encoding \((d \parallel d^*)\) and \((d \parallel d^*)\) for each \(d\)).